PROJECT REPORT

Survey of Gaelic Corpus Technology

Michael Bauer
Prof Roibeard Ó Maolalaigh
Rob Wherrett

for

Bòrd na Gàidhlig
Darach
Fèith nan Clach
Inbhir Nis
IV2 7PA

October 2009

Executive Summary

This report is compiled in response to the requirement from Bòrd na Gàidhlig (BnG) to investigate the current and future landscape of Corpus and related Language Technologies for Gaelic. The intention is to inform the strategic thinking about how and where to direct resources in the future to improve the technical aspects of the language development.

It has been based on a tripartite approach looking at:

- Current tools and technologies, their application and usability
- Wider ongoing and proposed developments - particularly amongst the other Celtic languages and opportunities for collaboration; and
- Aspirational needs of the user community, especially those professionally involved in the use of Gaelic through translation, education or media. The emphasis being on understanding how this translates into Speech and Language Technology (SALT) requirements.

The current focus of Gaelic SALT is seriously mismatched with where it should be, focusing on dictionaries and word lists rather than a comprehensive and integrated approach to codification and standardisation, essential SALT, corpus and lexicographical tools. These latter should feed into the everyday usage and uptake of the language in all its forms.

Furthermore the structures currently in place are disparate and uncoordinated. Best practice shows that there ought to be a formal Gaelic Academy that owns the codification (orthography, grammar and terminology) and is final arbiter on matters of technical aspects relating to the formal language. Other key aspects include the setting up of proper funding structures to give foundation to the technical work and to underpin the longer-term progression.

The **value to the language** of some of the proposals is significant (considerably in excess of £175m) provided that the recommended structured approaches are followed. Until now this fact has been signally obscured by the inability of virtually all stakeholders to see the hidden impacts of the current plethora of uncoordinated tools and approaches. Using best practice estimating models from the world of management consulting, this report shows how these hidden impacts can be given a capital value in the current Gaelic environment. When compared to the actual costs of getting it right - they also show just how much can be achieved.

Conclusion
Gaelic needs to revisit the structures in place to control and manage the technical aspects of corpus development and SALT. This in turn will lead to developing a completely new set of arrangements for moving forward.

Current developments need to be critically reviewed and in many cases suspended or cancelled altogether. Failures of Quality Assurance to date have wasted significant amounts of resource that could have been better utilised. Overall this highlights the need for a strategic/planned approach, for language professionals to be put in charge of corpus development and SALT; and to move away the carrying out of corpus development and SALT from the schools' sector.

The following immediate steps are to be encouraged:

- Core codification of the language as soon as is practically possible
- Transferring ownership of the orthography to the beginnings of a Gaelic Academy, making the latter responsible for the Quality Assurance of the codification and terminology development
- Development of a gold standard corpus, managed by a Celtic/Gaelic HEI. (This alone will have a capital value to the Gaelic Language community in excess of £173m and can be delivered for a tiny fraction of that cost.)
- Setting up a Governance Framework that has adequate guaranteed funding to enable the baselining of the Language and Tools
- Temporary suspension of the publication of authorised publications prescribing orthography/lexical/grammar usage (e.g. An Seotal, Ainmean-àite na h-Alba). This does not mean these project should cease all work, rather that they will be engaged in coordinating with the new Gaelic Academy training in best practice other foundation activity. Once the fundamentals (including core codification) are in place, these projects should then resume.

  If possible, delay the launch of new proofing tools until codification is complete and can be implemented.
- Beyond that adherence to International Protocols and Standards must be achieved at all levels.

In addition to these core recommendations, there are various additional recommendations, suggestions and ideas which can be found in the different sections of this report.

Collaboration with the following projects and partners should also commence at the earliest opportunity:

- Canolfan Bedwyr (Welsh SALT centre)
- Fiontar (Irish SALT centre)
- Foras na Gaeilge (Irish cross-border development agency)
- NCI (New Corpus for Ireland project)
- Professor Scannell (Professor of Computing, University of St. Louis, Missouri)
- Traslán (Translation company working in conjunction with Foras na Gaeilge)

Training and development of core teams must take place to engage in the technical work of Terminology Standardisation, and Corpus work

Finally wider aspects such as engaging with younger people to develop appropriate supportive technologies in areas like games and texting should be encouraged. All of this downstream activity should be professionally project-managed to ensure that there is cohesion and optimisation of resources and sharing of the collective output.

This represents a major challenge but if the proposals are followed Gaelic will take its place among the leading minority languages, rather than fire-fighting. The outcome can do nothing but good for the wider development and uptake of the language in the 21st century.

University of Glasgow

September 2009

CONTENTS

Glossaries

This is a glossary of technical terms and abbreviations used in this report:

Abbreviation	Description
CALL	**Computer Assisted Language Learning** refers to computer-based tools that support the teaching and learning of languages.
CAT	**Computer Assisted Translation** refers to translation done with the aid of tools such as translation memories (q.v.).
Corpus	The term **corpus** refers to collection of texts (and more recently, also recordings of a language). It can also, used loosely, refer to the totality of a language's inventory of words.
Diphone	**Diphone** is a term used in speech synthesis. It refers to two "sounds" next to each other, e.g. /ku/, /sa/, /kʲi:/. Diphone systems were amongst the earliest speech synthesis systems developed.
Generator	A **generator** does the reverse job of a lemmatiser. It takes a root (for example, *rach*) and generates the associated forms (*chaidh, thèid, tèid, dol,* etc).
Goidelic	This refers to the closely related group of Celtic languages including Manx, Irish and Scottish Gaelic.
GOC	**Scottish Gaelic Orthographic Conventions**, a rules framework for Gaelic spelling.
Lemmatiser	A **lemmatiser** is a tool that recognises different forms of a word. It would, for example, recognise *chaidh, thèid, tèid, dol, dhol,* etc as being derived from *rach*.
NLP	In this context, NLP stands for **Natural Language Processing** and refers to an area of linguistics that deals with processing human language on computers.
OCR	**Optical Character Recognition** is software that converts a digital image of text into text.
Open Source	**Open Source** refers to software where the code is freely available.
Part of Speech	**Part of Speech** (PoS) is a term for the various categories of words a language may have, such as nouns, verbs, adjectives, etc. They are also called lexical classes or categories.
Proofing tools	**Proofing tools** support users in producing and checking digital documents. They include tools such as spell-checkers, grammar-checkers and style-checkers.
SALT	**Speech and Language Technology** refers to the broad area where language and speech connect and includes both common use tools such as predictive texting and spell-checking and more specific tools such as lemmatisers and generators.
STT	**Speech to Text** refers to software that converts spoken language to text.
Tags	Linguistic corpora are often **tagged**. This means that material in the corpus has data tags attached to individual words that identify the nature of the word, e.g. *cù* could be tagged as noun/irregular/masculine. In modern corpora, such tags are not solely restriced to to parts of speech and can include tagging for phonetic features, intonation, etc.
Termbase	A **Termbase** is a database of (technical) terms.

Abbreviation	Description
TM	A **Translation Memory** is a tool that remembers previously translated strings of words and suggests these when new, similar strings are encountered in a digital document.
TTS	**Text to Speech** refers to software that converts a piece of digital text into spoken language.
Unit Selection	**Unit Selection** is a form of speech synthesis that produces very natural sounding synthetic speech. It relies on a specifically designed spoken corpus that has been phonetically transcribed. In unit selection synthesis, the engine always tries to locate the longest matching sequence. Although somewhat slower than diphone systems, unit selection is more natural sounding as it generally works with "bigger chunks of actual speech".

Glossary of key institutions and other bodies frequently referred to in this document:

Name	Short Description
Bwrdd yr Iaith Gymraeg	The Welsh Language Board (referred to as Bwrdd in this report).
Canolfan Bedwyr	A Welsh language technology centre at the University of Bangor, North Wales.
Euskaltzaindia	The Royal Academy of the Basque Language, amongst other things the regulator of orthographical and grammatical standardisation.
Fiontar	A language technology centre at Dublin City University.
Foras na Gaeilge	The cross-border development agency for the Irish language.
HPS	The Department for Language Policy of the Autonomous Government of the Basque Country.

1 Introduction

This report is compiled in response to the requirement from Bòrd na Gàidhlig (BnG) to investigate the current and future landscape of Corpus and related Language Technologies for Gaelic. The analysis and recommendations may be viewed by some as hard-hitting or critical at times. However, the intention is to give a best-practice and objective view to inform the strategic thinking about how and where to direct resources in the future. This will improve technical aspects of the language development.

It has been based on a tripartite approach looking at:

- Current tools and technologies, their application and usability
- Wider ongoing and proposed developments - particularly amongst the other Celtic languages and opportunities for collaboration; and
- Aspirational needs of the user community, especially those professionally involved in the use of Gaelic through translation, education or media. The emphasis being on understanding how this translates into Speech and Language Technology (SALT) requirements.

Overall the work was carried out using a combination of desk research, stakeholder reviews (including a wide-ranging online survey and a series of face-to-face workshops) and field research with other minoritised language organisations and academic bodies. This consisted of reviewing research papers, operational reports and meetings with representatives in Scotland, Wales and Ireland (including Northern Ireland).

The routes taken by a very wide range of minority languages have been considered, ranging from Northern Sámi to sub-regional dialects of Catalan. However, the emphasis has been on the Celtic language group due to their overall similarities and Basque. The latter is probably one of the best-organised in terms of current structures and has a similar linguistic distance to its neighbouring Spanish/French. One thing that is abundantly clear is the extent to which Gaelic lags seriously behind even some of the smallest European minoritised languages.

Aside from the linguistic aspects there has also been input from a business management perspective to ensure that proposals also meet quality criteria in terms of governance and ongoing management.

The intention here is to take a journey through the ideal world, compare that to what is currently going on with Gaelic and then to construct a Roadmap to get to the best possible future in terms of the SALT landscape. In so doing the report will address the needs for:

- A comprehensive list of existing tools, technologies and services
- An assessment of the value/effectiveness of each of the above - usually by reference to the impacts on the language environment and professional users
- A consolidated view of current and proposed developments and their likely value to Gaelic Corpus Planning. This will include structures and governance as well as technical matters.
- A consolidated Needs Analysis looking to the future. This will be based on a common group of common tools, technologies and structures that are deemed essential for any minoritised language; plus a practical view derived from the ultimate user of language technologies (be they professional or otherwise).

The highlighted gaps will be used to plan the way forward, taking account of the value chain. In terms of Corpus/SALT this is particularly important in delivering certainty in relation to areas such as orthography, grammar and lexicography. These in turn feed into the everyday usage and uptake of the language in all its forms.

Throughout the focus is on what constitutes a Quality Approach, especially since experience elsewhere has shown that it has been the lack of quality thinking at key stages that has resulted in massive rework or overhead in other areas to correct the consequential errors. Gaelic can ill-afford to repeat errors (of its own making or by other languages) if it is to catch up with where it needs to be.

Overall the aim is to deliver a comprehensive and comprehensible view of what is required. This is not an academic research report in the traditional sense (highly theoretical, extensive bibliography, based largely on academic input). Instead it follows a pragmatic approach based on research into experience in other languages. It aims to show the logical progression of technologies and language governance to deliver a stable, first-class language environment for Gaelic.

1.1 Structure and Explanations

The main report consists of the following subsections:

- A condensed summation of the approaches investigated in other countries (Section 2 - A General Schema for SALT Development)
- An overview of the current state of development of Gaelic speech and language technology (Section 3 - The Current Gaelic World)
- A description of the views and aspirations of Gaelic speakers and users that were collected during this research project (Section 4 - Aspirations of Gaelic Users)
- A detailed outline of the steps necessary for developing Gaelic corpus and speech and language technology (Section 5 - A Roadmap for Gaelic)

This is followed by a conclusion. Appendices contain the majority of the detailed source material, additional analysis and other data: the Index of Deliverables, Investigated Projects, Results of the Survey and the Outcomes of the Creative Workshops.

The following should also be noted:

- Scottish Gaelic is referred to throughout as Gaelic; Irish (Gaelic) as Irish; and Manx Gaelic as Manx.
- Institutions and other bodies are referred to by their native names and are shown in normal type e.g. Euskaltzaindia. Explanations of titles are given in the glossary on page 8 where appropriate.
- Key documents and reports from other sources that are available digitally are included in the Attached Files and referred to as necessary throughout the document.
- Commentary on various aspects of analysis is contained within text boxes.
- URLs have been included
    - Where it is felt that the subject matter would otherwise be difficult to find; or
    - Where the reader may want to find additional information about a body, institution or project.
- Attachments are supplied as separate digital files.

1.2 Value Estimates

Where estimates have been put on the value of components, this has been done by an accredited practitioner using the Monte Carlo Estimating function of Dimension Four ®. This is the world's leading Project Delivery toolset that focuses on where the value chain lies in organising change.

Monte Carlo Estimating within Dimension Four ®[1]

It is a common problem for projects not to be able to envisage the value of components, especially where they don't have an explicit price or value attached (unlike, for example, the case for a software development). The Monte Carlo Estimating method allows the reasonably assessable components to be identified - be they costs or impacts. The method is particularly good at putting values on hidden benefits.

For each component a low and high estimation is made taking into account real things that are likely to be occurring and for which it is possible to attribute some financial or numerical value. This is validated by common sense and knowledge of the context. By combining these it becomes relatively easy to arrive at an overall value of the impact. The results tend to be fairly accurate when viewed in retrospect - and are far more accurate than trying to assess values in isolation.

It becomes easy to identify a **single value** that is a pretty accurate representation of the likely final outcome by taking the median between the Low and High values that are produced from the sum of the components. In addition where values are recurring annually it is helpful to capitalise these to an NPV (Net Present Value). Practically this can be very complex but the nearest quick route is to take 10 years' worth of the value (i.e. Annual Value x 10) as being close enough for planning purposes.

For the purposes of this report - values attributed to benefits or costs using the Monte Carlo Method are shown in this way - i.e. Annual Median x 10.

The outcome is a useful benchmark of the values of benefits or avoided costs that can be achieved by a capital investment in change such as developing Codification or building a new Corpus Tool (see comment below on Speech Synthesis).

We have used this method to examine indirect benefits that are otherwise immeasurable. Hence, when we have tried to look at the Opportunity Cost of Speech Synthesis (see 5.4.6) it has been very easy to illustrate the benefits to the Gaelic language of having such a technology by looking at the costs of implementing the same effects via other means - assuming that were physically possible.

[1] Developed by Isochron Ltd in Edinburgh and now used by PricewaterhouseCoopers LLP internationally. See www.isochron.co.uk and www.pwc.com.

Monte Carlo Estimating has been used in an increasing number of Public Sector settings where items have been difficult to quantify (such as patient benefits in the NHS). The benefits to the research in the context of this report are to highlight the otherwise hidden values and costs that can be achieved or mitigated by implementing the structured approach we suggest.

In a language context this method of estimating has been used by the University of Edinburgh's **Bilingualism Matters**[2] Project to assess the potential positive economic impact on Scottish Society of bilingual children, as part of the developing business case.

[2] Professor Antonella Sorace Laurea MA PhD FRSA FRSE, Professor of Developmental Linguistics.

2 A General Schema for SALT Development

In the developed world, SALT permeates our everyday lives. Minoritised languages that do not manage to provide a reasonable and sustainable offering of SALT to their speaker base find it increasingly difficult to encourage and ensure continued language use in new domains, in particular amongst young people.

For most major world languages such as English, Spanish, Mandarin or Japanese, the development of SALT occurs naturally over time and in a widely dispersed manner. On the whole, resources are not an issue for such languages. They boast major research facilities, both in academia and the private sector; and products that are developed enjoy a large market. Also there are no questions over the actual survival of the language.

For lesser- or under-resourced languages such as Gaelic this is not the case. In many instances there are fundamental questions over language survival, resources are much less readily available, especially in the private sector and development rarely happens naturally.

Against this backdrop, maximising the use of available resources in the development of SALT for an under-resourced language such as Gaelic is paramount to achieving maximum benefits. To achieve this an overall strategy is needed both to minimise duplication, to avoid known mistakes and to decide on the most efficient ways of achieving SALT aims.

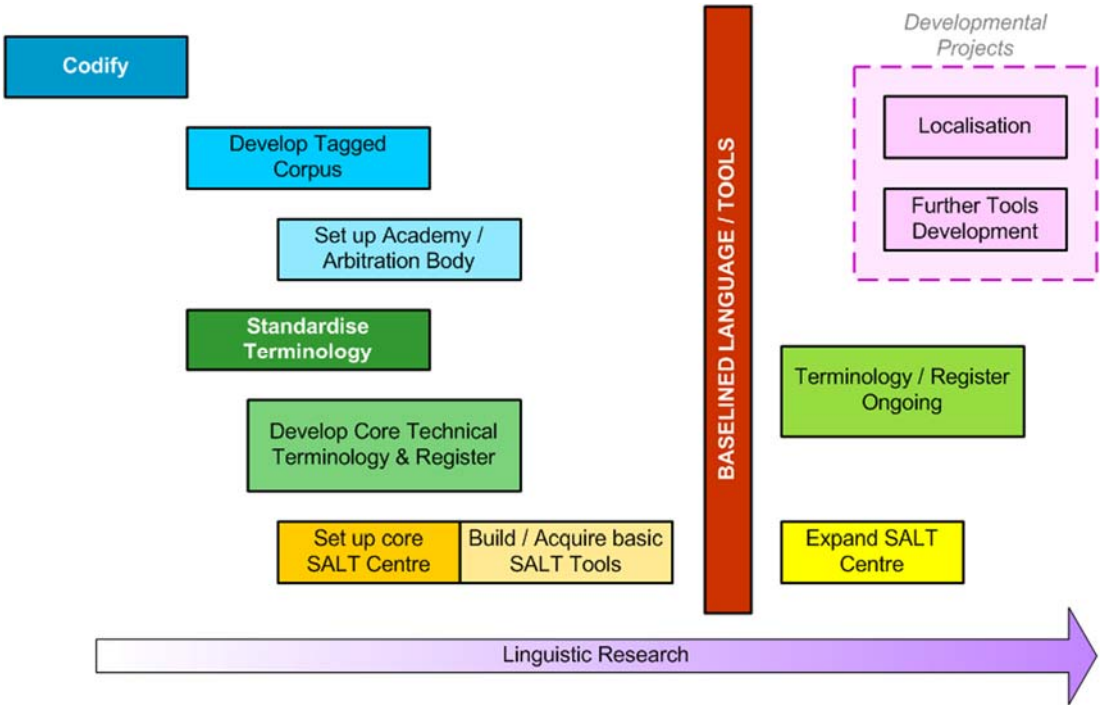


Figure 1 General Schema (ordered left to right)

2.1 Foundations

There are three broad aspects that form the foundation to successful SALT development:

2.1.1 Professionalism

Professionalism is vital, especially for lesser-resourced languages. For a variety of reasons, many languages initially do not have the required skills base within their communities and outside expertise is often both needed and desirable. If long-term measures to train the indigenous skills base are implemented early, the amount of outside expertise can be reduced over time without affecting the quality of the output adversely.

Opting for cheaper, less qualified personnel and/or preferring indigenous personnel without the necessary expertise from the start will ultimately lead to low quality outcomes, unnecessary expenditure and will not develop and increase the indigenous skills base.

2.1.2 Centres of Excellence

Dedicated independent centres of excellence, both SALT-specific and academic, foster research into fundamental linguistic issues, the development of tools and an indigenous skills base.

Such centres, where a critical mass of language experts and developers of technology are concentrated in a single space, have proven to be an extremely efficient way of responding appropriately and flexibly to the changing needs of a speaker community. They also function as breeding grounds for young talent (speaking the target language), a crucial task that continuous outsourcing of SALT contracts generally does not address.

Centres of excellence also allow long-term maintenance of technology related projects and skills. This is a key function. It is difficult to gain consumers of technology in minoritised languages as users in the respective language. On the other hand they are easy to lose to mainstream language technology if the technology is not kept up-to-date.

2.1.3 Collaboration and Funding

Collaboration is a vital tool in the development of SALT for lesser-resourced languages. Many tools share a common framework, in particular if the languages they were developed for are linguistically close. Pooling resources in collaborative projects reduces duplication, waste and ultimately leads to better tools for all parties concerned.

Collaboration also extends to the “availability” of research and basic tools. The more basic tools (such as lexical databases and software tools) are freely available (e.g. Open Source), the more likely it is they will encourage additional spin-off projects.

2.1.4 International Standard Protocols

Standards are vital in several senses. At the base of successful development lies a comprehensive agreed, stable and implemented formal language standard.

Beyond linguistic standards, adherence to international standards and best practice is also vital; particularly since under-resourced languages do not have the same capacity to develop their own quality standards to an equivalent level.

Standards are also crucial when filling positions. Although the long-term goal should be the development of an indigenous pool of experts, it is important that the technical skills needed for fulfilling a task are the overriding factor in employment questions.

2.1.5 Future-proofing

We live in an age of rapid technological development and changes. Adherence to international protocols that have developed best practice models and end-to-end thinking in general will produce outcomes that can be adapted to changes in technology with a minimum of effort.

2.2 Standardisation and Development

There are two vital aspects to standardisation, the development of a standard formal language and the development of terminology.

Unless an immense amount of re-working is considered acceptable, a standard spelling and grammar are vital within the realms of formal language. Once agreed, this needs to be implemented carefully but consistently across formal domains. This does not and should not infringe on the everyday usage, in particular of the spoken language.

Timeline of main standardisation events:

1891	Faroese
1928	Welsh
1945-58	Irish
1968-79	Basque
1979	Northern Sámi
1982	Romansh
1996	Friulian
2001/06	Sardinian

The later comprehensive codification and its implementation occur, the more likely SALT development will be held up at a fundamental level.

Beyond the development of a written standard, modern terminology must be streamlined, maintained, developed and disseminated to enable language users to function efficiently in 21st-century settings. Although competition in most fields is a healthy thing, in particular for smaller languages, competition in the field of new terminology development is extremely harmful and best carried out by a single body of experts. There are best-practice models and international standards that serve to facilitate the development of new terminology in an efficient manner.

- Irish terminology standardisation assigned to An Coiste Téarmaíochta in 1968.
- Welsh terminology standardisation taken on by the Canolfan Bedwyr from 1998 onwards.
- Nascent unit of Bwrdd yr Iaith Gymraeg from 1998 onwards (the current Terminology Standardisation and Translation Unit)
- Basque terminology standardisation assigned to the Terminologia Batzordea in 2002.

2.3 Development of a Corpus

Although a linguistic tool at first sight, a solid corpus is the foundation for a vast array of common-use tools today. It not only enables linguists to carry out groundwork research into linguistic issues that are needed by developers of SALT but it also facilitates the rapid development of up-to-date terminology resources and future, more advanced linguistic research.

Corpora come in various forms, from simple collections of texts to large databases that are tagged for a multitude of features. Untagged corpora are of little use in the development of SALT so the development of a so-called “gold standard corpus” (a manually tagged corpus) that enables the automated tagging of the remainder of a (sizeable) corpus is essential. Once set up, this corpus then needs to be maintained properly on an ongoing basis.

Timeline of publicly available tagged corpora of modern European minoritised languages:[3]

1994	Welsh	(Cronfa Electrone.g. o Gymraeg), 1 million words
2002	Basque	(Euskararen Corpusa), 4.6 million words
2002	Gaelic	(Will Lamb, private corpus): 82,000 words
2003	Frisian	(De taaldatabank Nijfrysk): 25 million words
2003	Irish	(Corpas Naisiúnta na Gaeilge): 8.5 million words
2006	Sámi	(Interaktiivalaš Korpus): 500,000 words
2006	Frisian	(Korpus Sprutsen Frysk): 650,000 words
2012	Irish	(Nua-chorpas na hÉireann): 30 million words

2.4 Developing Terminology Resources

Terminology resources support both expert and everyday users of the language in their use of the language in a variety of settings.

Traditionally the focus has been on printed dictionaries but in particular within the realm of lesser-resourced languages, there has been a strong shift to online and mobile resources as they reduce production costs, are immediately available to a large number of people and can be updated readily. These developments include both general dictionary resources and more specialised termbases containing technical terminology.

The shift to digital databases, such as lexical databases and termbases, also enables the development of advanced terminology resources such as TMs, proofing tools, assistive technologies, etc.

2.5 Developing Tools

The development of practical tools for users is paramount, as lesser-resourced languages rarely have the luxury of being able to develop tools on a whim. Such tools are varied and range from the basic to the extremely sophisticated. Overall there is a strong shift, in particular with lesser-used languages, towards using open-source software as the money saved in acquisition can be used for development. This benefits both the development of the indigenous skills base and the economic value and perceived value of the language.

Overall, those tools should be given priority that will benefit as high a number of users as possible and to the greatest extent.

The development of SALT in other countries shows that this is most effectively done through the establishment of centres of excellence that specifically aim to develop SALT resources for a language. Current developments include the move towards interdisciplinary technology networks that is a general feature of Research & Development in many fields.

[3] Details of corpora relevant to Scottish Gaelic corpus development are listed in Appendix 2.

Centres (with date of establishment) and examples of output and/or participation to date:

- Elhuyar (Basque, 1972): spell-checker, machine translation, research
- UZEI (Basque, 1977): terminology database, terminology implementation information system, corpus, lexical database
- Ixa (Basque, 1987): research, lexical database, spell-checker, lemmatiser, morphological analyser, electronic dictionaries
- Canolfan Bedwyr (Welsh, 1993): spell-checker, grammar-checker, speech synthesis, terminology standardisation software, lexical databases
- Fiontar (Irish, 1993): terminology databases, place-names database
- Euskara Institutua (Basque, 1996): research, corpora, speech synthesis
- Traslán (Irish, 2004): machine translation, translation memories, terminology database

2.5.1 General Tools

These are general tools to facilitate the use of the language in conjunction with technology. This may include hardware (such as specialised keyboards) or software tools such as on-screen keyboards or keyboard layouts. Such tools can be vital for the use of the language on computers, phones, etc.

2.5.2 General Software

General software enables users to work within the environment of their own language. The software needs for different users vary but initially, the majority needs should be given priority.

Starting dates of **Mozilla/Firefox** localisation projects that have produced releases for general use:

2002 Basque, Breton, Galician, Sorbian, Welsh
2003 Asturian, Catalan
2005 Irish
2007 Tatar
2008 Friulian, Occitan
2009 Romansh, Frisian

Starting dates for **Internet Explorer** releases:

1997 Basque, Catalan

Safari has no releases in minoritised languages.

2.5.3 Proofing Tools

Proofing tools such as spell-checkers and the more advanced grammar-checkers and style-checkers benefit both everyday users and specialists (such as translators). Provided their inherent shortcomings are borne in mind, they can have a great impact on the overall quality of digital or digital-derived output.

Timeline of proofing tools

1993	Welsh (Cysill), spelling and grammar
1998	Basque (Xuxen), spelling
1999	Frisian (Staveringskontrolearder), spelling
2000	Irish (GaelSpell), spelling
2002	Gaelic (GaidhealSpell), spelling
2003	Irish (An Gramadóir), grammar
2006	Gaelic (An Dearbhair), spelling

2.5.4 Speech Technology

Speech technology, both speech synthesis and speech recognition, have a wide range of applications. Both rely on more fundamental resources such as linguistic research into a language's phonology and intonation.

In mainstream languages speech synthesis is used in a variety of settings, most commonly in telecommunications and services. Another obvious application of speech synthesis technology (the "easier" of the two), is within education. This includes adult and continuing education.

Timeline of (first) synthetic Voice in:

2003	Catalan
2004	Welsh
2005	Irish
2007	Galician
2008	Basque

2.5.5 Computer-assisted Translation (CAT)

Irrespective of philosophical issues surrounding translation in smaller languages, translation is an integral part of most European languages. Training aside, tools that support translators can aid both the quantity and the quality of the output.

Basic CAT tools such as translation memory (TM) software generally exist but may need some language-specific development. They also suffer to some extent from their price versus usability and hence Open Source/free versions need encouragement. A lack of general information on TM software within the translator community is also a common problem.

More sophisticated CAT tools such as machine translation (MT) are currently less suitable for the Gaelic context. The development of MT relies on linguistic groundwork currently not available in Gaelic, such as bilingual corpora. It is also best suited to large translation projects using repetitive, technical language.

There is no general timeline for development of any of these sophisticated tools across minoritised languages. Development tends to be demand-driven.

2.6 Other Aspects

There are other aspects that need to be borne in mind.

2.6.1 Information Management

Information needs to be considered in two ways: within the speaker community; and towards (mostly) non-speakers.

Within the speaker community, channels must be established that encourage the flow in information regarding new developments, tools and projects, including possibilities for the exchange of information. This connection with the wider community is crucial as tools do not only rely solely on their pure existence but also information about their availability, their uses and limitations and the willingness to accept them. Aside from encouraging a sense of participation two-way communication flow also enables the relevant bodies to spot gaps, new possibilities, developments and potential needs more quickly. For the most part this is easily achieved by using (e-)newsletters, online fora, blogs and similar technology.

The flow of information towards non-speakers is an important mechanism to enable non-speakers use the language more or more efficiently. The majority of European minoritised languages run help lines (phone, fax, email, text) that deal with small enquiries regarding language use, such as spelling, grammar and short translations. These services are geared towards public sector employees but most readily accept enquiries from a wider audience.

- Berripapera[4]: (e-)newsletter of new developments in Basque services, resources, etc produced by the Department for Language Policy
- Freagra: Irish hotline, includes translation/spelling support
- LinkLine: Welsh hotline, includes translation/spelling support
- SALTcymru: (e-)newsletter for Welsh SALT-related news

2.6.2 Research

Contemporary linguistic research into the language is also an important aspect that often requires encouragement. Lesser-resourced languages often suffer from a lack of research into aspects such as (contemporary) phonology, grammar, syntax, semantics, etc. This frequently hampers the development of more sophisticated tools such as grammar-checkers as developers normally rely on existing research.

Various methods are employed to encourage this type of research, most commonly in specific research centres or by encouraging, for example, under- and post-graduate research grant schemes and existing academic institutions to conduct the necessary research.

Examples of institutions for research or the support and promotion of linguistic and scientific research:

1918	Eusko Ikaskuntza (Basque)
1940	School of Celtic Studies at Dublin Institute of Advanced Studies (Irish)
1972	Institiúid Teangeolaíochta Éireann (Irish), closed in 2003
2007	Innobasque (Basque)
2009	SALTcymru (Welsh)

[4] See Supplemental Files for a sample.

3 The Current Gaelic World

The GAP analysis of existing Gaelic-related projects shows that for the most part, Gaelic does not score well when viewed within the context of other lesser-resourced languages.

Other languages on the whole follow best-practice models and international standards and focus their efforts on developing centres of excellence, local skills bases and building solid foundations. In stark contrast, Gaelic related efforts (with a few notable exceptions such as the SMO's Pools-T project or the interuniversity project Faclair na Gàidhlig) happen in isolation without laying proper foundations, and tend not to follow best-practice models nor adopt international standards.

For example, in Wales, an early dictionary project led to the creation of an industry standard lexical database by a dedicated team. This in turn led to the development of a spell-checker, followed by a grammar-checker and facilitated the development of various other tools. In the process, a centre of excellence was established that continues to develop new tools and fosters an indigenous skills base.

By contrast, the main Gaelic spell-checker was outsourced to an English-based group and created as a flat Word document. Although the group in question has been able to utilise the file for some other developments, this has not facilitated the development of a centre of excellence, an indigenous skills base or outside follow-on projects. Had a lexical database been created, for example, the later creation of a word prediction tool would not have required the creation of a new text file.

3.1 Stakeholder research

Apart from investigating methods and technologies, views from Gaelic speakers and users were also sought. The main input here was via 5 creative workshops (Glasgow, Edinburgh, Inverness, SMO, Stornoway) and an online survey. Although reponses to the survey were invited from as wide an audience as possible, the respondents to the survey represented a somewhat skewed sample.

Overall there were 108 returns to the survey, of which a subset (19) were translators who answered some supplementary questions. By comparison with similar surveys into other minoritised languages, this was a fairly robust response.[5]

For example, 20% of respondents identified themselves as translators. Since it would have been impossible to achieve a structured cross-section of the Gaelic speaking community the results of the survey must be interpreted as representing the professional end of the Gaelic community. Nevertheless, it is this professional end that highlights the lack of information, low uptake of tools and problems with terminology and standardisation. If professional users are confused, what chance is there for the lay individual?

3.2 Foundations

3.2.1 Standards

On the whole, standards are under-developed or not adhered to (see 3.3).

3.2.2 Centres of Excellence

None currently exist in the sense described under 2.1.2. There is a considerable pool of skills both in terms of linguistic and technical skills in Scotland and the wider Goidelic arena but for the most part these are isolated individuals (both academic and otherwise). If they work on Gaelic-related projects, in many cases this tends to happen in their spare time.

[5] For example, the survey for the 2008 SALTcymru report had 48 respondents.

As a result, the development of an indigenous skills base is lagging as most SALT related projects are outsourced to various individuals and groups.

3.2.3 Collaboration

Most collaboration with other language groups is restricted to individual agents networking with relevant groups and people in Ireland, Wales, etc.

Even within Scotland, little co-ordination or collaboration takes place even between larger and related projects. For example, there was virtually no contact and collaboration between the OpenOffice localisation project run by LTS/Cànan and the Windows Vista and MSOffice localisation projects run by TELI/Microsoft. Also the terminology developed for both has, to date, not been made public.

3.3 Standardisation and Development

Some work has been carried out on producing a standardised spelling (GOC). The survey of the GOC framework (conducted as part of this research) shows that it is not seen as comprehensive by the majority of users and that dissemination and acceptance are lacking. Little has been done to date to produce a research-based standard grammar for the formal language.

There is currently no governing body for standardisation although SQA, as the body responsible for GOC, is currently the de-facto governing body on Gaelic spelling. Control of standardisation by such an "outside" body which does not answer to the community is a situation not commonly found outside Scotland. Instead, control is usually exercised by independent bodies of experts, such as the Académie Française, to ensure independent and expert arbitration on the topic.

In terms of international standards in terminology work, these are neither known nor adhered to by most bodies involved in terminology work. The majority of work in this field is carried out without the input of trained terminologists or the use of terminology development management systems.

3.4 Development of a Corpus

No gold-standard tagged corpus exists. The closest is a dormant project by Will Lamb who produced a tagged corpus of spoken and written Gaelic. That corpus contains c.80,000 words, but this is not publicly available and is currently dormant.

Contributing to the Faclair na Gàidhlig project, the Digital Archive of Scottish Gaelic (DASG; see Appendix 2) has begun the first stage of working towards a corpus, Corpas na Gàidhlig, covering the historical period up until the 21st century. This Corpas na Gàidhlig project is currently working on digitising c.220 texts from this period but has not yet started work on adapting or designing a corpus engine. DASG draws on expertise from the SCOTS project (see Appendix 2 for details).

DASG is in contact with the Gaelic digitisation project by the National Library of Scotland which, amongst others could provide sources of data for a future corpus.

All other corpus projects that contain Gaelic are little more than small collections of digital text (LER-BIML, Tobar na Gaedhilge, etc.).

3.5 Developing Terminology Resources

Existing terminology resources present a slightly more varied picture. Although a number of SALT projects such as spell-checkers and word predictors currently exist, none of these have led to the creation of lexical databases or similar resources. Overall, co-ordination between ongoing projects is low and resources that are developed within these projects are not made available to a wider audience.

3.5.1 Dictionaries

Traditional printed dictionaries are predominantly Gaelic to English,[6] of which Colin Mark's Gaelic-English dictionary was published in 2004 and Dwelly's in 1901. The remaining printed sources broadly fall into the category of pocket-dictionaries.

There are three main sources of terminology online:

- An Stòr-dàta (SD)
- Dwelly-d; and
- Faclair na Pàrlamaid.

Of these, only the SD and the Faclair na Pàrlamaid contain modern terms. Although the SD at its heart has the printed 1994 edition, the online version is a mostly un-edited word-list. Yet according to the survey, the SD and Dwelly-d are the two most frequently used and appreciated resources.

Currently only the SD and Dwelly-d are interlinked, all other digital resources (online and offline) have to be consulted separately, which is a major inconvenience for everyday and professional users.

The follow-on project from Dwelly-d, Am Faclair Beag (AFB), is working on the production of a new dictionary resource including information on sounds, grammar and other information alongside Dwelly-d content. It aims to integrate various terminology resources and to date, the 23,000 entries from Faclair nan Gnàthasan-cainnte have been added to it. It also has a new utility that allows registered native speakers to vote on their familarity with a term, thereby conducting a linguistic "audit". To ease the workload, AFB contains a word-form generator that generates verb, adjective and noun forms that can be predicted by rules. This enables editors to automatically generate forms which are then checked manually. It is currently the only dictionary project that is known to be actively working on a lexical database. The framework for AFB has been largely created but new content creation is slow due to limitations of time. None of the projects (SD, Dwelly-d and AFB) have a specific plan or budget for future developments and all rely on voluntary input.

Wordlink, a prototype of a browser-based tool that links words in a web-page to a dictionary, was developed as part of the European Pools and Pools-T project. The work on Wordlink is carried out by SMO staff and funded by European money.

A Gaelic thesaurus, based loosely on the spell-checker, has been produced by TELI for LTS and is due for publication in 2009. The output is envisaged to be a digital online file.

3.5.2 Historical Dictionaries

The Department of Celtic at the University of Glasgow collected a substantial amount of material in fieldwork from 1966-1996. It contains questionnaires, word-lists, recordings and other materials from a wide range of dialects, including material from less well researched areas such as Nova Scotia and Kintyre. This was originally to be the basis of the Historical Dictionary of Scottish Gaelic (HDSG). However, work on the dictionary was brought to an end in 1996. It has been replaced by the inter-university project Faclair na Gàidhlig (See Appendix 2). The materials collected for HDSG are being digitised by the DASG project (See Appendix 2).

[6] Excepting MacLennan's bidirectional dictionary.

The Faclair na Gàidhlig project, a Scottish inter-university project was set up in 2003 with the aim of producing a historical dictionary for Gaelic and a number of spin-off outcomes which will contribute to Gaelic corpus planning. It will contain material from the HDSG fieldwork but will be primarily based on a full-text database consisting of c.220 texts selected from the entire historical period of Scottish Gaelic up until the 21st century. Work on digitising the texts has already begun at the University of Glasgow under the auspices of the DASG project.

Although most language communities at some point embark on such historical projects, the value of purely historical dictionaries in the development of SALT is limited. For instance, the only practical tool (apart from the dictionary itself) that the Frisian historical dictionary project yielded was a spell-checker. (This despite the fact that it had been lauded as a major development.)

3.5.3 Terminology Development

An Seotal, set up in 2007 as part of Stòrlann, is currently the only dedicated terminology project. Its database currently contains c.500 terms focussing on scientific and mathematical terminology and is in need of expansion. Although the aim is to have established 1,500 terms by the end of 2009, the project has no specific output targets. Originally set up to run until 2008, it has currently been extended until 2011.

It uses in-house staff (a translator and a dedicated project officer, but neither a terminologist nor a lexicographer) to audit and create terminology. The terminology is then passed by volunteer teacher panels, with an advisory panel (meeting once every 2 months) that has the final say on contentious issues.

We recognise the constraints under which this organisation is currently working - meeting the regular demand for new educational publications alongside its other development work.

Taking the wider view, the project could benefit from membership in national or international terminology associations[7] and training with regard to international standards on terminology development.

There is wide use of non-standard grammatical terms in their material, both English and Gaelic (such as *suidheachadh ainmneach* vs *an tuiseal ainmeach, possessive case/suidheachadh ceangailte* vs *genitive case/an tuiseal ginideach,* etc. The meaning of these new terms is explained on their website, but it is not clear what the intended benefit of creating such new grammatical terminology is.

Of the small number of specialised dictionaries such as Faclan Ùra (SRG) or the Maths Glossary (Stòrlann), few are available online. In the case of Faclan Ùra, the printed version was not widely available to the general public. None of the terminology lists said to be circulating within the GME system are available outside the education system.

The place-names project, Ainmean-àite na h-Alba (AÀA), currently has a part-time staff of three (1.9FTE) with funding in place until 2011 to research authoritative forms of place-names for bilingual signage. As such its main emphasis has been researching place-names along certain routes without specific goals regarding place-names outside these routes. The first part of the project has been devoted to developing a database framework, in spite of the fact that virtually identical projects exist in Northern Ireland and the Republic of Ireland whose technology could likely have been shared.

[7] Such as the European Association for Terminology for example (http://www.eaft-aet.net/)

The database currently contains c.100 place-names (of 1600 that have been researched since April 2007) and is not accessible to the general public, although it is the stated aim to do so in the future. Some of the data researched is available as PDF files but worryingly contains errors. Beyond 2011 there are currently no plans in place for this project and there are no plans to include data previously researched, such as the place-names lists collated by Iain Mac an Tàilleir for the Scottish Parliament.

There is no project that deals with other onomastic issues such as names and surnames.
As a daily broadcaster of Gaelic content, BBC Alba is heavily involved both in terminology creation and dissemination. Within the time constraints of the broadcasting schedule, staff consult a variety of existing terminology sources and discuss terminology with colleagues to determine and create appropriate terms.

Although BBC Alba and MG ALBA are represented on Bòrd na Gàidhlig's Resources, Terminology and Translation Committee, there is no dedicated terminology team within the BBC dealing with terminology. Companies providing content for programmes occasionally provide lists of terminology used/created for programmes but there seems to be no overall policy on terminology. A small amount of the terminology used eventually finds its way to a small online wordlist on the BBC website[8] but the majority of the terminology chosen or created is not held centrally or distributed effectively between the different branches of the BBC dealing with Gaelic programming.

3.6 Developing Tools

There are a number of tools that have been developed for Gaelic but again, on the whole, development is characterised by isolated projects that rarely foster future development in line with best-practice in other countries.

As regards software in general, there is a strong movement toward Open Source software globally. This type of software development is mostly driven by developers in mainstream languages and as a result, a myriad of applications have been or are being developed, ranging from games to desktop publishing software and operating systems. For a variety of reasons such as cost and the freedom to adapt (and localise) Open Source software, such software is increasingly becoming mainstream around the world and found from government offices and private corporations to schools and private homes. In the survey, 30.5% of respondents reported that at least some Open Source software was already currently used at their workplace.

This affords an unprecedented opportunity for lesser-resourced languages like Gaelic. The use and development of such technology can facilitate a much wider provision of tools to everyday users and professionals at little cost. Funds that would otherwise be spent on licensing, usually outside the indigenous community, can be used to further develop such tools, thus benefiting the local skills base and the wider economic situation of language users and professionals. The development, use and promotion of such Open Source software should therefore be a priority.

3.6.1 General Tools

Very little has been done in a structured way in this area. Users of the language are left to their own devices in finding ways of tackling Gaelic on technology platforms. Although relatively convenient ways of dealing with the accented characters exist (such as using the Irish keyboard settings), a substantial number of users are not aware of these methods or they lack the computing skills to enable them.

[8] Facail Fheumail http://www.bbc.co.uk/scotland/alba/naidheachdan/facail/

Other general tools, such as word predictors, even such developed in conjunction with major Scottish bodies, are rarely advertised and promoted sufficiently and as a result, uptake is low. For example, the word predictor developed by Penfriend in conjunction with Stòrlann and LTS, is virtually unknown in the wider community.

3.6.2 General Software

General software such as operating systems and office applications are currently substandard. OpenOffice 1.1 is the only Gaelic office application currently available. (Note - the current version in general use elsewhere is v3.2) There is no operating system and the only web-browser available is Opera 6.05 (current version 10.X).

Localisation projects for Microsoft Office, Windows Vista (produced by TELI/Microsoft) and OpenOffice 3.2 (LTS/Cànan) are complete but have not been released to date. As previously, this was issued as contract work with no long-term maintenance provision and thus also fails to meet long term goals. The update of OpenOffice 3.2 was a rare example of a TM (Poedit, an Open Source application) being used. There are no concrete plans to make the TM available, although Cànan might consider doing so.

Few other software localisation projects exist. The vast majority of them are Open Source applications such as forum software, Ubuntu and Firefox or community translation projects such as Google where the translation work is carried out by volunteers. This entails problems such as slow progress, variations of spelling, style and terminology as well as problems with general language skills.

3.6.3 Proofing Tools

The only proofing tools currently available are spell-checkers. Of these, only two can be described as functional, one for Windows (An Dearbhair) and one for MacOSX (GaidhealSpell). The latter is little known. Usage of spell-checkers is low overall, even amongst translators. Of the latter group 59.9% of respondents stated they never or rarely used one. Common grievances were problems with the installation and/or running of the software, a lack of updates and a lack of trust in the programme.

Unlike in most other countries where spell-checkers are derived from lexical databases of some sort, the Dearbhair was compiled as a Word document. While this enabled the development of the spell-checker, it does not facilitate the development of additional tools nor easy maintenance of the existing tool. GaidhealSpell is a corpus-derived tool and therefore operates on a different level and requires an improved corpus for future development.

3.6.4 Speech Technology

Speech technology is virtually non-existent. There have been several small pilot projects in the past to develop a so-called diphone system but no functional tool to date has become available. No attempt has been made to date to produce a TTS system based on current state of the art systems that, in any case, have moved away from pure diphone systems.

The closest to a functional system that exists is the Irish Cabóigín (see Abair in Appendix 2) project that has always harboured the intention of including Gaelic at a future date. The current Donegal Voice is marginally capable of Gaelic TTS as well as adequate Irish TTS.

3.6.5 Computer-assisted Translation (CAT)

There are currently no specific Gaelic CAT tools, either in the shape of available translation memories (TMs) or TM software and there are no other tools such as machine translation (MT) available. The uptake of TM software, which can be used for most language pairs irrespective of the software interface language, is virtually nil. The reasons given for this low uptake are a lack of information and the price of proprietary software.

In the light of the project survey responses, wider research into the topic amongst freelance translators, long-term goals in Gaelic SALT and the wider international view, the development of a single and affordable TM system based on Open Source software should be a priority. Thus the recent decision by UHI to adopt a proprietary system for internal use, with a view to future licensing outside UHI are not likely to benefit the wider Gaelic translation sector.

Although ultimately feasible, MT currently is a tool that requires much foundation work. It also is a technology that needs to be "handled with care", is generally used by professionals in translation and overall, expectations of quality usually exceed the output. Such a system should therefore **not** be a priority in the current Gaelic context. Any future centre of excellence dealing with Gaelic SALT should, however, keep abreast with new developments in the field and react appropriately if the situation changes. If approached, MT development should be based on leading edge technology current at the time.

3.6.6 Information Management

Dissemination of information in either direction and to both speakers and non-speakers is haphazard at best. There are few regular channels of information that focus on developments in the Gaelic world overall and none that focus on SALT. Except for a site maintained by SMO that aims to collect a simple list of available Gaelic sites and tools, there is no resource on- or offline that specifically aims to inform everyday users on the availability of tools or that provides support.

3.6.7 Research

The main thrust of academic research into Gaelic has traditionally been aimed at history, literature, philology and dialect studies although valuable work has been carried out in the field of sociolinguistics in more recent times. Several comprehensive descriptions of various dialects exist. In terms of Gaelic linguistics the main focus has been into Gaelic phonology more than anything else. Some recent research has begun to investigate advanced aspects of Gaelic grammar, syntax and semantics.

But overall, technical linguistic aspects of Gaelic remain under-researched. There are no formal initiatives currently that specifically foster Gaelic-related research projects by under- or post-graduate students of non-Celtic/Gaelic subjects in Scotland, for example, in IT or linguistics departments.

3.6.8 Non-Gaelic Skills Pool

It is worthwhile noting that while there are few Gaelic-related projects at Scottish universities outside the Celtic departments, there are various academic departments that have skills and experience that may be of use in future Gaelic projects (see Appendix 2).

3.7 Planned Projects

There are currently few planned developments in the area of Gaelic SALT.

3.7.1 Faclair Bun-tùsach

There is a proposal by TELI, a Sussex-based group of non-Gaelic speakers, to produce a concise English-Gaelic dictionary containing some 50,000 headwords. While overall the development of English to Gaelic lexicographical resources is desirable, within the overall strategy for SALT and corpus planning, this project falls short on a number of points.

Funding external groups outside Scotland does not contribute in a major way to developing a local centre of excellence nor does it foster an indigenous skills base. TELI also has a questionable track record in producing lexicographical products that facilitate the future development of additional tools as they appear to have a preference for compiling flat Word documents rather than databases for example.

A major lexicographical endeavour such as the compilation of a 50,000 headword dictionary must be undertaken in view of the overall development aims for Gaelic. It should conform to international standards in lexicography and facilitate future follow-on products. For example, the compilation of such a dictionary in an appropriate database format would enable the development of a lemmatiser (see 4.1.12).

3.7.2 Grammar-checker

There are currently no ongoing projects developing a grammar-checker. Although no concrete plans currently exist, James Galbraith (Edinburgh) is investigating the possibility of developing a grammar-checker for Gaelic

3.7.3 Talking Dictionary

There is a (currently unofficial) proposal to develop a talking dictionary. Marc Farr of the North Highland College has already indicated that he is interested in using the Am Faclair Beag framework, as it already exists.

3.7.4 Grammar Dictionary

There is a proposal by James Gregor (Morar) to produce an online Gaelic grammar dictionary based on existing publications to explain basic and advanced topics of Gaelic grammar.

Although a single, well-built online resource would have some benefits, there are currently various websites and publications that already deal with explaining the basic issues of Gaelic grammar. The problem with advanced grammatical issues is that they are under-researched and not sufficiently well-described in the literature. Until such research has been carried out, such a project is unlikely to yield any immediate additional benefits. It may be worthwhile, however, to consider converting some of the existing grammars that have been produced with public funding into online resources.

Once grammatical standardisation and more linguistic research into advanced topics has taken place, there is a definite need for the production of grammatical resources. This applies to users at all levels, but especially topics that are not covered in detail to date.

4 Aspirations of Gaelic Users

Demand for a wide variety of tools and mechanisms were found in the course of the research, in particular through the creative workshops[9]. For some, demand was almost universal.

4.1.1 Communication, Consultation and Information

A fundamental problem with a lack of communication and consultation was mentioned and criticised at virtually every workshop and by numerous respondents in the survey. Participants felt that, in general, the views of the wider community were neither solicited nor their wishes integrated sufficiently. The lack of communication between "Gaelic bodies" in general and the community was also lamented strongly. As a consequence various respondents expressed their (pleasant) surprise at the high level of consultation with the community in this research project.

Modern technology could easily be used to address both issues. Communication channels such as (regular) e-newsletters, blogs, webcasts, feeds, fora and online survey tools are relatively cheap and could go a long way to enabling the wider community to make their views heard.

There was also a very strong demand for the immediate launch of a site (new or connected to an existing site) that brings together information and help on all currently existing Gaelic tools. Many tools were unknown to participants who commented on the fact that they had only found out about their existence via the online survey. For example, the existence of a Gaelic interface for digital whiteboards is virtually unknown in- or outside the GME sector.

Such an information site should deal not only with advanced issues but also extremely basic ones that affect the use of the language in technology in everyday life. For example, guidance on the best ways of entering accented characters, changing document language and how to disable automatic word correction in word processors were mentioned.

Existing tools should be made available to more/all users of the language, including those outside the GME sector, including material such as the Guthan nan Eilean[10] project.

There were also calls for an information service that could provide Gaelic support. Such a service could provide short translations and language support, especially to the public sector. Such a service should be accessible via various channels such as emails or text. As the geographical location of such a service is irrelevant, it could be used to provide Gaelic-related employment in remote areas, provided competent staff can be found and/or trained.

4.1.2 Gaelic Academy

Calls for a permanent and independent Gaelic Academy (that is, independent of the education and government) also were made by a large number of people, in line with other European languages. It was envisaged that this would:

- Be staffed/operated by a mixture of linguists, professional users of the language, native speakers and other experts. Educationalists' role should be limited to advice on matters such as practicality in a teaching context.
- Deal with completing the task of orthographical standardisation.
- Work towards producing a standard form of the language (grammar, style) for use in formal contexts.

[9] Some of the views, suggestions and requests that came out of the workshops were not strictly related to SALT. As they provide a source of wider input from everyday and professional users of the language, a full report for each workshop can be found in Appendix 4.

[10] See www.smo.uhi.ac.uk/smo/naidheachd/fiosan/guthan-nan-eilean.html

- Be open, inclusive and pragmatic in its approach and operation.
- Be the governing/regulating authority for linguistic issues related to the language
- Either work towards the standardisation and development of specialised terminology or be the ultimate authority on terminological work carried out by another body.

4.1.3 Codification

Most respondents felt that the current degree of confusion and variation on orthography and grammar was unhelpful at best. The following points were raised and advocated:

- The need for a comprehensive rule set to be worked out by an academy (see 4.1.2), based on the groundwork laid by GOC but capable of considering "all options".[11]
- Once a comprehensive rule set is agreed, there will be the need for stability of those rules over a period.
- From the outset, every care must be taken to ensure the formal, standardised language is presented in context. It must not be seen by the community to be in conflict with the richness and diversity of everyday language and its dialects, spoken or written. Proper guidance must be issued and disseminated on the intended usage of the standard.

4.1.4 Standardisation

It was also stated by most people that in relation to formal uses of the language, the current degree of variation of technical terminology was unacceptable. The following points were raised and advocated:

- Re-invigorating the use of dormant native terminology in preference to new coinages.
- The need for a considerable team to deal with terminology standardisation in the first place, development in the second. This should be done in an open and transparent manner and be in conjunction with the terminological tools described in 4.1.5. The only current initiative, An Seotal, was criticised for seemingly occurring in isolation, the low amount of output to date and lack of features (such as reference to register, examples of usage, etc). It was described as "yet another wordlist" and confidence in the tool is low.

Standardisation also should be considered outside the narrow range of terminology. Other communities increasingly use online repositories of standard forms, templates, phrases, etc to reduce (at least in theory) the amount of variation between translations of the same material. This is particularly common and desirable within the public sector.

4.1.5 Terminological Tools

Also virtually universal were calls for a single, "all singing, all dancing" online resource for terminology. It was repeatedly stated that the current setup where sources, especially of technical terminology, are spread across more than a dozen different locations is not acceptable. This is particularly true for academics, translators, teachers and other professional users of the language for whom the current situation results in huge losses of time. Such a resource would:

[11] In the sense that the reversal of GOC changes must be an option if expert opinion deems this necessary.

- Be available online, accessible via mobile technology and digitally off-line with the possibility of regular updates. Provision of terminology in PDF/Word (and similar) text based formats was deemed extremely unhelpful, in particular for use on increasingly common mobile technologies.
- Strive towards collating all terminology available to date in one place.
- Include both technical and non-technical terminology, including place-names, abbreviations, names of official institutions, etc.
- Include grammatical information, examples of use, information on regional terms.
- Contain guidance on pronunciation (including sound), dealing both with mid-ground pronunciation models and diverging dialectal forms. In line with Gaelic phonology, this should exist not purely at the single word level but include longer strings (e.g. article + noun).
- Maintained and expanded with time.
- Future-proof terminology resources by using industry-standard forms of data storage and annotation.

As a first step, it was suggested that the existing features of the online Stòr-dàta should be improved.

It was also criticised that to date, virtually all lexicographical work focussed on the formal language and that, in particular for learners, there was no material available to deal with highly informal registers (such as slang dictionaries).

4.1.6 SALT in the Wider Context

More sharing and cooperation in the development of tools and technologies, in particular amongst the Celtic nations was advocated by various people. It was pointed out that, through collaboration, both the target market of any such developments would increase and provide "better funding for better tools", possibly even the development of some commercial products.

Most importantly though it was criticised that in terms of technology, Celtic languages, for the most part, are always engaged in a game of catch-up and are rarely at the forefront of development. It was suggested that by setting up centres of excellence for research and development in collaboration with other Celtic languages, this could be addressed. The stated aim of such a centre or centres would be the development of new/unique technologies that primarily would be made available to small languages only (giving them an edge over mainstream languages).

It was also mentioned in this context that the development of services only available in the target language should also be an important factor to increase the usage value of the language. For example, programs without subtitles on BBC Alba were mentioned in the workshops as a factor that increased the actual and perceived usefulness of the language.

Progress in conferencing software and virtual reality in conjunction with measures to promote, teach and increase use of the language could have the added benefit of providing employment opportunities for Gaelic speakers in remote areas via the web. However, it was pointed out in this context that such developments would require considerable technical support and careful liaison with communities as there is a certain perceived reluctance in relation to the usage of new technologies.

This "reluctance" is a factor that needs to be borne in mind in the development of any new technologies from the outset. In order to increase acceptance, technology must be made easily available, easy to use and explained and supported properly, especially when using Gaelic as the meta-language. Consultation and early testing by everyday users must be made a priority in the development of such tools.

4.1.7 SALT Tools

- Predictive texting was seen as an essential tool, in particular for encouraging use amongst young people.
- A grammar-checker, even a simple one, was suggested. Proofing tools overall should be designed to be intelligent (e.g. suggesting guidance when similar errors are repeatedly made).
- Integrating proofing tools, including dictionaries and thesauri, into MSOffice/ OpenOffice.
- Overall, SALT such as TTS or STT has to be carefully designed and must be functional at the point of release to the public. To date, the experience with such technologies has not been very positive and releasing not fully-functional tools in the Gaelic context may have an extremely detrimental effect on confidence and usage. Such negative experiences include the non-Gaelic sector where complaints about the inability of such systems to deal with Scottish accents or place-names were common.

  Developing STT to support translators in transcribing large volumes of spoken material was mentioned as being potentially useful by translators at the workshops.

  CAT tools should be designed so they can be easily updated and maintained. TMs should be made available to translators, and if possible, enable data collection to update existing termbases.
- Development of lemmatisers and generators to enable better linking between digital documents and online dictionaries. This would, for example, enable linking forms like *mòra, mhòra, mhòir*, etc to a dictionary entry for *mòr*.

4.1.8 SALT Technology in Education

Apart from general educational problems such as a mismatch of expectations during the transition from secondary to tertiary education, it was felt that at this level Gaelic speaking students were being "lost in the crowd", with the main focus being on the easily identifiable students of Celtic/Gaelic subjects. The following ideas were floated:

- The use of (networking) technology, in particular through university intranets and during the registration process. This would improve the participation of Gaelic-speaking students of non-Gaelic subjects at non-Gaelic HEIs in Gaelic events, projects and campaigns from the outset. Nonetheless, more traditional approaches such as use of Gaelic language officers may also offer possibilities for promoting such links.
- Currently Gaelic does not have the linguistic registers, precise terminology nor indeed habituation in most subjects to even consider teaching through the medium of Gaelic at tertiary level. To encourage grassroots development of such registers, terminology and skills, it was felt that technology could be used to provide baseline training.

  By using technology such as conferencing, a minimal but regular service could be provided to (for example) Gaelic-speaking students of biology to gradually develop their personal language skills and the register. At some universities such "study groups" exist at an informal level but it was felt that by formalising them, providing technical and educational support and extending it to all Scottish universities through the use of technology (such as online conferencing), the pool of interested students could be extended significantly.

In this context the potential of technologies such as MIT Open Courseware[12] was also mentioned. This could add both to the academic development of the scattered Gaelic-speaking student population and, through taster options, provide a more realistic insight into tertiary level education for students in secondary education.

- The use of technology to create a "Blasroom" to help learners acquire better pronunciation, to expose them to more native Gaelic speech and help them communicate with native speakers more effectively. In this context the idea of a speech-based virtual reality environment was also suggested (for details see Appendix 4 - Glasgow Workshop).

4.1.9 Mobile Technology

Greater use of mobile technology was suggested. One suggested application was the use of GPS with a Gaelic information service (see 5.3.4) that would provide people with Gaelic related information such as the pronunciation and meaning of Gaelic mountain-names or Gaelic language services (such as shops or accommodation) in the area. The latter could also be used to enhance the economic value of Gaelic in the community.

4.1.10 Translation

The use of technology in translation was advocated for a number of purposes:

- Providing better information for Gaelic translators in terms of working methods, training opportunities, available tools, etc. A specialist Gaelic translators' forum was mentioned in this context, as well as the need for a professional body and professional validation. Such a body could provide and maintain these services.
- Providing (better) guidance to HR/translation agencies regarding Gaelic translation.
- The use of conferencing technology to provide training opportunities for Gaelic translators. As most Gaelic translators have existing work commitments and frequently work as part-time translators, training opportunities must take these limitations into account.
- MT was deemed to be potentially useful, if of sufficient quality.

4.1.11 Research

The point was made that across Scotland not enough use was made of under- and post-graduate students to advance research into various aspects of the language (not related to literature or historical linguistics). Greater promotion of Gaelic-language related research and research projects, both within and outwith the Celtic/Gaelic departments, is needed. This was felt to be particularly necessary in the field of ITC and linguistics. This use of students in research is commonplace in other languages and while it may require the input of a native speaker, it does not necessarily require the students to be fluent speakers.

4.1.12 Other Points

Other points that were raised are:

- Currently the only sizeable tools that are accessible English to Gaelic that deal with idiom are Faclair nan Gnàthasan-cainnte, Dwelly-d and and Roy Wentworth's Gaelic Words and Phrases from Wester Ross. It was felt too much emphasis is currently placed on grammatical correctness but that much more needed to be done to develop tools that support good idiom.

[12] A project where recordings of lectures and lecture related materials are made available internally to students.

- Views were expressed that the widespread reluctance to openly criticise and to admit to failure are a major obstacle in the development not only of Gaelic technology but the wider Gaelic context.

  It was suggested that, in the context of SALT, intelligent proofing tools and other similar tools (e.g. an “Idiom of the Day” application) could be used to depersonalise criticism of linguistic errors and improve language skills.

5 A Roadmap for Gaelic

The following roadmap, the result of a GAP analysis between the general schema and the current Gaelic world, details the necessary steps to achieve the ultimate goal of a comprehensive formal language standard and a range of SALT necessary for users of the language in a 21st-century information society. It represents the most direct route to arrive at the level of development found in other, more advanced, European minoritised languages such as Basque, Irish and Welsh.

There are broadly-speaking two stages to the development of Gaelic SALT. Stage 1, the Linguistic Foundation, deals with underlying frameworks, issues and resources that facilitate the development of the more sophisticated tools of Stage 2.

The vast majority of this roadmap for the development of the Gaelic SALT sector coincides both with current and historical developments in other European minority languages but also more theoretic models that describe the staged developmental needs of any language within the SALT context. This includes, for example, the concept of BLARKS (Basic Language Resource Kits) modelled by ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association) or the roadmap developed by IXA at the University of the Basque Country (see Appendix 2).

All these identify broadly the same (basic) stages and tools: codification of orthography and grammar, a written/spoken/bilingual tagged corpus, terminology standardisation, basic SALT specific tools (lemmatisers, analysers, parsers, etc). These are then followed by more sophisticated tools such as speech synthesis.

5.1 Governance Framework

Governance in Language Development - particularly in the arena of SALT - needs to be clearly defined. There are specialist and general functions. In addition experience in other languages shows that getting this right is crucial. As mentioned elsewhere the reliance on amateur governance (in a linguistic context) is not acceptable. Matters need to be placed in the hands of recognised specialists - albeit they may report on progress to the more generally-based BnG in the latter's role as custodian.

In our considered view, this structure should look like Figure 1, with overall Funding controlled from the top and distributed following the solid lines:

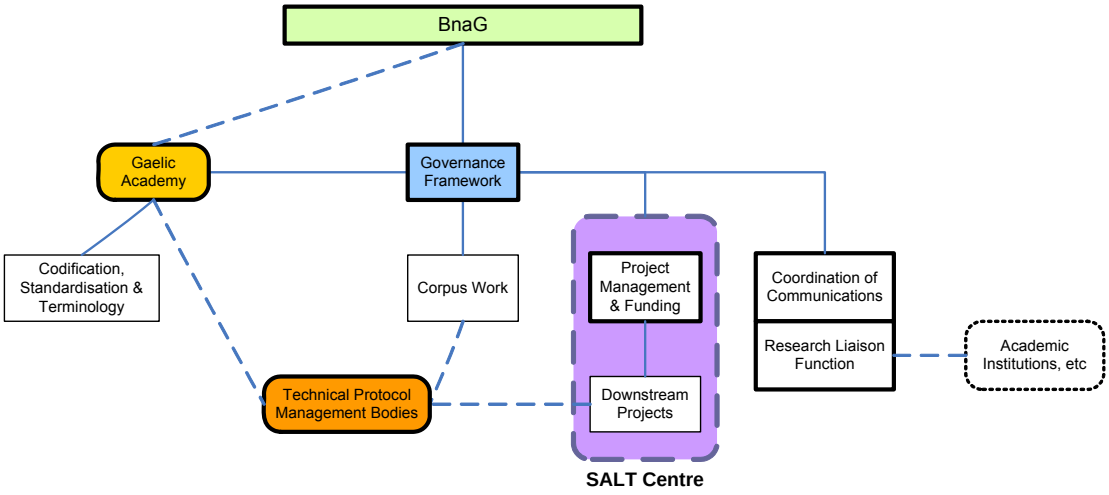


Figure 2 Governance Structure

The Governance Framework as shown may be a subsidiary function of BnG - there is no prescriptive definition. It will, however, require to be routinely administered.

5.1.1 Strategic Business Case

To move the whole framework forward will no doubt require the construction of a Strategic Business Case for presentation to the Scottish Government. This will follow directly from the issuance of a Strategic Plan for Gaelic SALT (**not** the existing Language Plan) that sets out the roadmap. This plan will be based to a greater or lesser extent on the recommendations set out in this Report, taking into consideration any political or other factors.

In pure management terms it is not feasible that the overall exercise be carried out without such a business case since this will provide the rationale for funding the proposed Governance Structures, some of which may need political approval.

Public Sector mandatory requirements for Programmes (and it must be clear at this point that the SALT and related developments are going to require a major effort) must gather approval from the Funding Body before commissioning. Whilst we cannot say that following MSP (Managing Successful Programmes) will be the "right" answer in this instance as a control mechanism, there is mandatory requirement from the Office of Government Commerce (OGC) which is part of the UK Treasury, for clear and accountable management at every stage. This mandate is equally applicable in Scotland, irrespective of other devolved matters.

5.1.2 Funding for Framework

The Plan for Gaelic will set out proposed Frameworks and Governance and no doubt some work on those can press ahead regardless of other matters. However, the whole structure will need to have agreed and stable funding in place. We envisage that the Funding for the Framework will cover the core activity and provide a platform for other bodies to develop additional funding for downstream activities and projects. Downstream funding may come from a multiplicity of sources and is not covered here.

5.2 General Principles

In order to achieve the desired state of Gaelic being fully equipped to deal with 21st-century concepts and the necessary technology to function in an ICT society, there are a number of specific steps required. Certain principles should be an overarching theme in all of these.

5.2.1 Detailed Principles

5.2.1.1 Openness and Professionalism

The vast majority of the required measures are aimed at the general Gaelic-speaking public. A culture of openness, inclusion and engagement with these vital stakeholders will improve acceptance and usefulness to the end user.

Professionalism must be encouraged. The development of an indigenous, Gaelic-speaking skills base is a long term goal that cannot be achieved overnight. As ensuring the quality of any output is paramount, it will therefore be necessary in a number of projects to bring in, even on a long-term basis, expertise from non-Gaelic speakers. If knowledge transfer schemes are built into such projects, these can contribute significantly to the training of a future Gaelic-speaking skills pool. Continuing the current model of outsourcing SALT related work to companies and groups outside the Gaelic community will perpetuate the lack of a Gaelic skills base.

To ensure the right skill set and mix to set up a Gaelic centre of excellence, collaborating with experts from existing Celtic centres of excellence who have taken this approach (Fiontar in Ireland and the Canolfan Bedwyr in Wales) will prove invaluable to ensure this is carried out to the highest possible standard.

5.2.1.2 *Centres of Excellence*

A Gaelic centre of excellence, a SALT centre, will meet strategic aims at several levels. It will provide a nucleus of experts that will support associated projects with technical expertise (such as adapting/building termbases and corpora), lay the groundwork for future developments, start developing an indigenous skills base and in general, provide the most direct route to a range of SALT tools.

By directly targeting the development of such a centre along the lines set out below, it will be possible to curtail the (longer) process through which such centres have evolved elsewhere.

5.2.1.3 *Collaboration and Funding*

Collaboration with existing projects, either locally in Scotland or internationally, must be a consideration in every Gaelic project. There is a large linguistic overlap, in particular with Goidelic languages that will allow sharing and easy adaptation of existing tools.

A much wider approach to funding Gaelic SALT and the associated research is required. There is no reason why Gaelic-related projects should rely solely on funding earmarked for Gaelic. Greater use should therefore be made of alternate funding schemes such as general research grants (e.g. the AHRC or Google Summer of Code [13]) or European regional funding, especially in cross-border projects.

Sharing of basic resources, for example, lexical databases or termbanks, must be promoted on a much wider basis. It should be made a funding requirement of future projects that, where applicable, such basic data must be made available for inclusion and sharing. Sharing of basic data will ultimately lead to a larger number of tools available to users.

5.2.1.4 *International Standard Protocols*

International standards must be used and adhered to in the development of the language and associated technologies as much as possible. This in itself will go a long way towards future-proofing developments and reducing the need for rework.

Membership or association with relevant professional bodies (such as the European Association for Terminology [14]) should be encouraged strongly to improve knowledge and compliance with protocols and familiarity with and input in new developments. Regular attendance at relevant international conferences should also be a regular activity.

5.2.1.5 *Future-proofing*

Future-proofing, in particular in technology developments, is vital for lesser-resourced languages. Generally they can ill-afford having to recreate tools due to a lack of future-proofing in the first place.

[13] Google Summer of Code http://code.google.com/soc/

[14] AET www.eaft-aet.net

Some functions in relation to future-proofing technology in a wider sense would also fall to the Bòrd itself. For example, raising awareness of the typographical needs of Gaelic amongst public (and private) sector technology procurement so that when new technology is acquired, it is made sure that it is at least theoretically capable of handling accented characters. Or encouraging greater use of the language on digital displays, a measure that is low-cost with a wide impact This is used much more widely in Wales, the Basque Country and to some extent Ireland than is currently the case in Scotland.

5.3 Stage 1 - Linguistic Foundation

This is shown in graphic form on the following page (Figure 3). Thereafter detailed notes explain each component and the relationships between them.

The diagram shows key dependencies between projects and structures. Importantly it also shows at high level the activities that will be required to enable them. The whole is based on understanding of best practice, not only in Language/SALT Development but also in management control. No other language has had the opportunity to step back and deliberately devise a process structure in this way and therefore other examples have complexities and structures that are in some cases unhelpful or even contradictory. For instance many have failed to consult on Best Practice in setting up Terminology Standardisation or have not been aware at outset of the need for some structures.

Gaelic needs to take advantage of this opportunity to get far and fast.

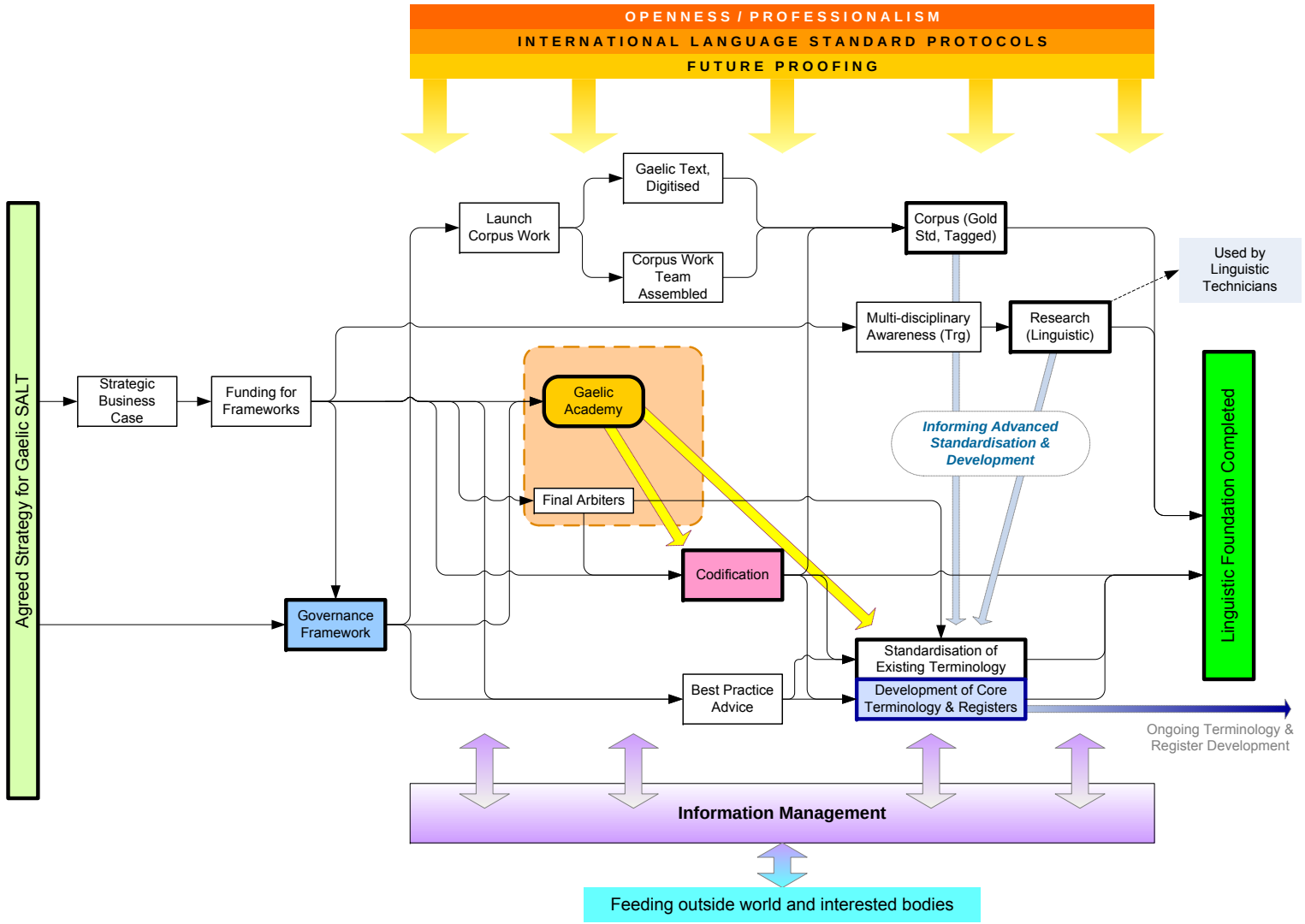


Figure 3 Linguistic Foundation

5.3.1 Standardisation

Three main steps are required as regards the standardisation of a formal register of Gaelic. It is likely that they will initially require larger temporary teams to deal with immediate codification and standardisation of existing terminology, followed by smaller long-term teams to provide continuous development of terminology and occasional arbitration on codification issues.

Codification is and has been an emotive issue and in most cases has taken decades if not centuries to evolve and settle. Smaller languages can rarely afford the luxury of taking a century to settle this fundamental issue without seriously affecting development. Development of a solid framework based on scientific method and principles, proper dissemination, adoption at the national/regional level and stability represent the "least painful" way forward.

The Value of Codification

Simply put, without Codification it is not possible to develop a consistent Standard High Register. That particularly impacts the translation/production of formal/professional documents and educational publications.

- To evaluate the impact we may suppose that currently there are 50 - 70 Public bodies regularly producing Gaelic Language documents each year.
- For those, the amount of documentation may be 2-5 documents each.
- Each document has an average length of 5,000 words.
- Translation costs a total of 15p per word (direct and indirect).
- Rework is going to be 40% of the base costs (even though currently rework may not happen it has a hidden effect of an equivalent value by spreading bad practice and adding to the confusion in the linguistic landscape). This is **real** damage even though it is invisible because ultimately more people get more things wrong without realising it. Unravelling the consequent mess at some future point will be very costly in terms of re-education and rewriting of important materials.
- For each formal document produced there are between 5-20 informal documents (teaching aids, local community leaflets, etc.) that base their standards indirectly on what they see and read in the formal literature.

The total annual hidden cost estimates are therefore based on:

{[No of Public Bodies] x **[No of Documents]** x **[Average Length]}** x **£0.15** x **40%** x **{No of other documents based on original}**

This is the amount of invisible damage being done to the language environment each year. We have capitalised this to give a resulting **NPV estimated at £11.9 million**

> **Note:** This takes no account of the **additional value** to technical development of items like spell-checkers. These are more likely to be developed commercially (and hence at nil cost to the centre) if there is an adequate rules base to work from.

Orthographic reforms that are not based on these principles will be problematic at best. The 1996 German spelling reform led to huge controversies and eventually required intervention from Government and a judgement by the High Court. This unsystematic approach to simplification has led to considerable costs in German-speaking countries over the years, both tangible and intangible. For example, rules on writing words together or apart[15] were reduced to 7 rules. However, this required knowledge of 253 usage conditions, 45 sub-rules, 2 specifications, 15 optional rules, 153 conditions and exceptions contained in 33 word-lists. The reforms have led to continued divisions between users of the old spelling (including some major publishers) and the new spelling to the detriment of all involved.

Gaelic spelling continues to be messy. This is particularly true as in lesser-resourced languages the turnover of new publications is low and new conventions are therefore slow to permeate the range of available books. Even users of GOC are often inconsistent in their use of rules and gaps in the framework are deal with differently by different users. This continued mottled picture is, however, not only a problem but also an opportunity, as thorough and comprehensive reform will hardly muddle the current waters any further and lead instead to a clearer situation.

5.3.1.1 *The "Gaelic Academy"*

The control of codification and terminology standardisation & development must be centrally organised by an independent body and its work based on the basic principles stated in 5.1. Such a body must be the final arbiter in codification and standardisation issues. It must also network and collaborate with similar bodies internationally.

As a matter of urgency, it needs to acquire ownership of the current orthographical framework (GOC), formal terminology development (An Seotal) and wherever possible ownership or agreements with existing terminology resources.

[15] This is the equivalent of the Scottish Gaelic *hyphen - no hyphen* question.

Timeline of official governing bodies:

1582 Italian, *Accademia della Crusca*
1635 French, *Académie française*
1713 Spanish, *Real Academia Española*
1906 Galician, *Real Academia Galega*
1911 Catalan, *Institut d'Estudis Catalans*
1919 Basque, *Euskaltzaindia*
1919 Irish, *Rannóg an Aistriúcháin* (orthography)
1938 West Frisian, *Fryske Academy*
1951 Sorbian, *Serbski Insitut*
1968 Irish, *An Coiste Téarmaíochta* (terminology)
1985 Faroese, *Føroyska Málnevndina*
2005 Cornish, *Keskowethyans an Taves Kernewek*
2006 Aragonese, *Academia de l'Aragonés*
2006 Kashubian, *Radzëzna Kaszëbsczégò Jãzëka*

Languages where an official orthography was introduced by state decree. In all cases there were several conflicting orthographies prior to standardisation (in the case of Sámi there were 9).

1979 Northern Sámi
1996 Friulian
2001/06 Sardinian

Even lesser-resourced languages often have a central body that has authority over issues to do with codification and terminology development. Codification (and its implementation) is such a basic issue underlying all other developments that, in most other language communities, it has already been **comprehensively addressed** during the 20th century.

Welsh orthography was largely standardised in 1928, with amendments in 1987 (both by committee). Irish was standardised between 1945-58 by the state translation office (Rannóg an Aistriúcháin) and Basque by the Euskaltzaindia between 1968-79.

Failures to do so or the continued presence of diverging orthographies, for example, in Sardinia (latest attempt in 2001/2006) or Romansh in Switzerland 16 , have presented significant obstacles not only in the development of SALT but also the wider language.

Similarly, cases where divisions are being actively created, such as the establishment of the 1998 Acadèmia Valenciana de la Llengua to promote the regional Valencian dialect as being separate from the over-regional Catalan are both divisive and counterproductive.

The adage that “many cooks spoil the broth” certainly holds true in terminology development. Either by design or by trial and error; other European minoritised languages have adopted a single independent authority approach or are in the process of doing so.[17]

[16] In the case of Romansh the 1982 standard orthography and grammar in particular this has led to massive savings as historically official materials, including educational publications, had to be produced in 5 dialect varieties. This approach was officially abandoned in 2003 in favour of the new standard.

[17] With a tendency for this being organised OUTSIDE the Education Sector and direct Governmental control.

5.3.1.2 Codification

Comprehensive codification of a **formal register** is a prerequisite for virtually any development of basic or advanced resources and technology. Failure to codify will result in significant loss of both time and money; or loss of quality for most SALT developments and users of the language.

The codification team needs to consist of, at minimum, the following roles:

- Gaelic linguists (2-3) with an in-depth understanding of high-register grammatical processes and forms of the language. In as much as data/ awareness of this exists, they also need good comprehension of recent trends and developments in the language.
- Linguists/Philologists/Dialectologists (1-2) with an in-depth understanding of the historical development of the language (including the links between Irish and Gaelic), the historical development of the orthography and dialectal variation.
- Gaelic phonologists (1-2) with an in-depth understanding of Gaelic text-to-phoneme rules.
- Expert native speaker users (2-3) of the language, such as experienced editors in Gaelic publishing, translators who have practical experience of gaps and problem areas in the current system.
- An expert in language policy and planning.

There will also be a requirement for User Acceptance Panels (UAP) to ensure the final framework is not wholly unacceptable to native speakers who use the written language formally and informally on a regular basis. This should include publishers, educationalists, translators, developers etc. It should also include input from the increasingly large and active fluent learner community. The UAPs should openly invite participation from a wide range of people to increase buy-in from the community.

For those experts in full-time employment it may be advisable to arrange buyouts/secondments to enable them to participate in an appropriate manner. In line with the general recommendations, a maximum of openness towards the wider community should be ensured.

As a matter of urgency, codification of basic orthographic and grammatical issues must be addressed. As the modern vernacular is under-researched, this will likely have to be based on existing information about conservative forms of the language. For example, much work involving native speaker judgement tests and the investigation of stress placement remains to be done on compound words in Gaelic. This work would help to establish clear rules for the use of hyphenation in Gaelic orthography.

Some aspects will require native speaker panels representing all major living dialects for judgement test. For example, the current free-for-all in terms of hyphen placement (supposed to indicate stress shift) can only be sorted by native speaker judgement tests.

Other languages have made the mistake of letting things progress at their own pace (resulting in long delays) or have gone for speed over quality. Neither has resulted in a first-time success story. We envisage that codification should be completed within approximately one year, subject to adequate Quality Assurance to ensure the standards are of the highest quality. This will ensure that speed of delivery does not compromise quality.

A key function of best practice for developments involving technical experts is the use of professional high-calibre project management of the work streams alongside very high quality subject-matter expertise. This approach will help to deliver the desired quality at an acceptable speed.

Ideally the existing framework (GOC) should be the basis of full codification. However, based on the maxims of **professionalism, scientific method** and **cohesive principles**, if expert opinion recommends divergence from this framework, their verdict must be accepted as an expert independent decision.

> Codification has surprisingly wide implications beyond the obvious.
>
> For example, the merger of *'nan* and *nan* under GOC results in a computational cost in speech synthesis because the words *'nan* and *nan* are treated differently phonologically.
>
> On the written page, a reader is capable of using wider context to distinguish the meaning (*in their* vs *of the*). For a computer, this is virtually impossible. This ambiguity would, in turn, result in less natural Gaelic.

Publications from authorised sources (such as Stòrlann) dealing with orthography/grammar will need to be reviewed within the light of codification. This is particularly needed as regards some of the unilateral modifications such as the use of dialects possessives (*ar h-, ur h-*) over traditional forms (*ar n-, ur n-*).[18]

The survey shows that at present confusion regarding orthography is widespread; therefore amendments to the framework would not significantly increase confusion, especially if implemented properly. However, the result would be a vastly improved basis for future work.

Following codification, this needs to be followed by a significant period of stability and implementation.

> - Orthographic issues and irregularities (e.g. capitalisation, st/sd in proper nouns, accents on inherently long vowels)
> - Hyphenation
> - Adaptation and transliteration of English/Latin/Greek loan-words (e.g. phonological issues, current violations of Gaelic letter-to-sound rules)
> - Treatment of non-Gaelic words in Gaelic texts (e.g. lenition or non-lenition of letters not in the Gaelic alphabet)
> - Treatment of acronyms, Gaelic and non-Gaelic
> - Unresolved grammatical issues and irregularities (e.g. compound nouns and long noun phrases, hyphenation, verbal forms)
> - Unresolved register issues (e.g. use of dative case, use of genitive case)

Some codification issues (both basic and advanced)[19] may need to be re-visited at a later point when research has been carried out and the Corpus has been set up to allow refinement of guidance on advanced issues of grammar, style and register.

[18] See Gràmar na Gàidhlig, (ISBN 086152 753 4) Stòrlann/Acair 2002, Section 1.7.1.

[19] For example the use of èa vs eu vs ia

- **Openness**
  Engagement and collaboration with professional users of the language and other interested parties in the process will increase acceptance and dissemination
- **Professionalism**
  Use of experts (professional users, linguistic experts) will lead to a clearer, more comprehensive framework
- **Standards and Future-proofing**
  Future-proofing should consider the implications of (creating) irregularities and ambiguities on the future development of SALT.

5.3.1.3 Terminology Standardisation and Development

Ultimately, control over all publicly-funded terminology projects must be brought together centrally at the Academy. This will not remove jobs from the various bodies but it will bring considerable benefits by reduction of complexity and rework.

Current standardisation and terminology development projects would benefit greatly from immediate upskilling. Experts (e.g. Fiontar or the Canolfan Bedwyr) should be brought in as a matter of urgency to help analyse and if necessary improve the framework and provide on-the-job training to involved staff. This leads us to consider the impact on An Seotal.

Taking account of their particular constraints we believe the way forward is to add some additional resource that will focus immediately on getting to grips with training/development. That would allow the current staff to continue to support their publications - albeit with an eye to consulting regularly in order to ensure a minimum of future rework. Over time the enlarged team could then progressively refocus on terminology standardisation and development.

As it is not feasible to halt development until basic codification is completed and training has been completed, it must be accepted that at some point terminology work carried out prior to a cut-off date will have to be revisited.

Once core codification of spelling and grammar has been addressed, standardisation of existing terminology must be addressed as a matter of urgency. Acceptable time-frames can be ascertained by collaborating with centres of excellence such as the Canolfan Bedwyr, which have carried out similar work previously.

A permanent, paid core team must be set up which, with the help of (outside) best practice advice on terminology standardisation and development, will develop the framework for standardisation and development. If at all possible any external training should be provided by qualified Irish experts due to the linguistic similarities between the languages which lead to similar needs, problems and solutions. Available professional technology such as Maes-T (see Appendix 2) must be brought in to streamline the workflow and move away from less productive approaches.

The use of a certain amount of voluntary work on terminology panels is inevitable but should be kept to a minimum. To overcome the problem of lack of sustainable progress, it would be helpful to set expectations as to the workload that will be involved and get agreement to deadlines from the outset.

Standardisation primarily needs to address the following:

- Technical terminology
- Place-names
- Names and abbreviations of official bodies
- Names and surnames

Once the proper framework, teams and resources have been put into place, the first priority should be the standardisation of existing core items. This task will include the extraction of terminology from existing sources such as the Microsoft localisation project, the OpenOffice project, Opera, Google and published (educational) resources. Once complete, work needs to commence on developing prioritised areas of terminology. Initially in terms of technology inevitably and unavoidably this will result in a skewed development of technical domains. Although GME needs are likely to feature high on the list of priorities, attention must also be paid to the needs of other sectors, especially ICT but also tertiary education and the public sector. Unresolved issues regarding orthography, transliteration, etc that surface in the process must be referred back to the codification team.

Place-names standardisation should also consider immediately what additional features (e.g. post-codes for settlements) should be added to make the resulting database as flexible as possible for future use in technology and not only as a “list of place-names”.

Within the above framework and until the establishment of an academy, a better staffed, resourced and trained team including An Seotal and Ainmean-àite na h-Alba could very possibly continue to carry out terminology standardisation and terminology development work. Once core areas of the other areas of standardisation have been addressed, priorities and formal goals for future development should be agreed. The output of place-names, surnames, terminology and official names must be held in the same location for end-user convenience.

Especially given the lack of widely read print media in Gaelic (in the form of newspapers, magazines, etc.), BBC Alba and Radio nan Gàidheal have a key role as disseminators of terminology. In particular BBC Alba is, de facto, also involved in terminology creation on a daily basis. While it may not be practical in the near future for the BBC to rely solely on outside sources for its terminology, BBC Alba shoud be

- Included in the development of structures that deal with terminology development and standardisation.
- Encouraged to develop a coherent policy on developing, collating, storing and making terminology accessible.
- Develop a policy for BBC Alba productions and client productions to ensure the use of Gaelic terminology wherever possible. The use of English terminology in new BBC Alba programming (for example the use of mostly English plant names in the recently broadcast series Fàs Slàn) should be discouraged. As virtually all programmes are subtitled, this would in no way hinder comprehension.
- Urged to accept the need to adhere to agreed terminology once this has been properly standardised.

What value for a Lexical Database?

This is going impact on the build of both general and specialised dictionaries. In addition the database will feed the development spell-checkers and more sophisticated grammar-checkers.

Given that these latter items are dealt with in the evaluation of the Corpus (see 5.3.2 below) and largely represent the proportion of those items that are not resulting from the Corpus, we make no attempt to put separate values on them here. However, it must be clear that a Lexical Database has a value to the language that is of similar order of magnitude to components derived from the Corpus.

- Represents 20-30% of the value of a Grammar Checker
- Represents 90-95% of the value of a Spell Checker
- Represents 30-70% of the value of a good Thesaurus
- Represents 80-95%+ of the value driven by a Lemmatiser/ Generator (which enables good look-up tools).

The focus of such terminology work must be on developing the **modern language**. Projects dealing with more historical aspects of the language are not usually considered high priority in the early stages of development. Notwithstanding, given the paucity of terminological resources in Gaelic and the lack of development of modern technical registers, it may be necessary to utilise sources from an age that in a mainstream language would not be considered contemporary.

- In the context of Standardisation & Development, **Openness** is going to require engagement and collaboration with professional users of the language and other interested parties. Containing the work within a confined group operating behind closed doors is not best practice.
- It is essential to engage with **Professionals** in this core work. Use of experts (trained terminologists, lexicographers, language experts) will ensure maximum quality of the output.
- Following international standards on terminology development will help future-proof the terminological aspects; use of the technical guidelines (for example, on design of lexical databases) will facilitate increasingly useful foundations for SALT development. This fulfils the needs for **Standards and Future-Proofing**.

Key features of the online terminology database:

- Contains standardised technical terminology, place-names, names and abbreviations of official bodies, surnames
- Will ultimately also contain common vocabulary
- Gives reliable information on pronunciation (including sound), grammatical information, examples of usage
- Online and downloadable

5.3.2 A Gaelic Corpus

An early start on developing a Gaelic corpus is necessary to facilitate research into the language. This will also aid the development of SALT resources in less time, at lower costs and, generally, more sophisticated tools. To this end, the corpus must contain a gold-standard tagged core and be of considerable size.

> A corpus with a gold-standard core is an internationally accepted concept of an ideal. It consists of:
>
> (i) A core set of data that has been manually tagged for the desired features; this will train an "automatic tagger" for additional material added to the corpus
>
> (ii) Additional data that has been tagged automatically

Association of such a project with a university department, as is commonplace amongst lesser-resourced languages, is likely to be a favoured setting for a Gaelic corpus project. As this project is likely to involve a considerable amount of digitisation, facilities such as book scanners commonly found in university libraries would also be accessible. There are potential employment opportunities via the internet for proofreaders in remote Gaelic speaking areas.

Collaboration with the Irish corpus (NCI) should seriously be considered, possibly even to the extent of "joining" the NCI project. Collaboration with the National Library of Scotland Gaelic digitisation project must also be investigated.

> Key features of a Gaelic corpus:
>
> - Corpus (with gold-standard core)
> - Tagged for parts of speech (additional tags to be determined)
> - Development of a lemmatiser/generator in conjunction with terminology developers
> - Capable of dealing with spelling variations as contradicting orthographies exist
> - Includes written, spoken and bilingual material with capacity to include audio material
> - Generous cut-off date due to relative paucity of very recent material

There are various possible sources for data that could feed a corpus. Overall, future funding for projects that involve recording of speakers and production of native written materials should automatically request permission to include materials into the corpus to pre-empt legal questions surrounding intellectual property.

Potential sources include:

- Public donations
- Publishers
- Media
- Gaelic bodies
- Public sector
- Existing corpora
- Academia

The Corpas na Gàidhlig, part of both the Faclair na Gàidhlig and DASG projects (see Appendix 2), is a possible starting point. However, as it currently contributes to an historical dictionary project aiming to cover the entire historical period of Gaelic, assurance is needed that in the early phases:

- the primary goals of development will focus on a sizeable corpus of modern materials, selected according to criteria normally applied in modern corpus design.
- a tagged, gold-standard core will be produced which can be used to tag additional data.

If such goals could be formally agreed, this could be an ideal setting for a Gaelic corpus.

The value of a good Corpus?

This facilitates 3 major components of the SALT:

- Grammar Checker (which has a value broadly equivalent to that of a Spell Checker); *by up to 70-80%*
- Better Dictionaries; *by up to 60-80%*
- Speech Synthesis; *by around 20-40%*

We can put values on these individual components as follows:

- Grammar Checker is worth **£18.88m** using an equivalent value to Spellchecking (see Spell Checker notes under 5.4.3.2) x 70-80%
- A good Dictionary is probably worth a similar amount to a spell-checker - say 80-100% of that figure. Giving a putative value of **£23.8m**
- Speech Synthesis (see 5.4.6) is evaluated at **£135m**

The total NPV estimates are therefore close to £173m:

Of course the actual costs of developing the Corpus are going to be a tiny fraction of that.

5.3.3 Academic Research

It is commonplace outside Scotland for research into linguistic aspects and the crossover between language and technology to be a multi-disciplinary affair, often not carried out by speakers of a particular language but by researchers with the help of native speaker informants.

In the Gaelic context, this type of cross-disciplinary collaboration is infrequent and largely informal. This kind of activity should be given much greater prominence. More academic research into Gaelic linguistic topics must be encouraged and supported as much SALT relies on its existence. This will also inform advanced standardisation issues of the language.

Existing Gaelic linguistic research projects, researchers and Celtic/Gaelic departments could usefully further develop their academic networks nationally and internationally. This includes university departments of linguistics and ICT. Ideally such networks should be coordinated via the SALT centre but as an immediate measure even a simple web-forum could be used to improve the situation.

This will promote more advanced research into linguistic aspects of the language (syntax, phonology, semantics, etc). Celtic/Gaelic departments need to further raise the importance of this type of research and increase possibilities of carrying out and participating in such research in collaboration with departments outside Celtic/Gaelic. In this context the Bòrd should also set up a graduate funding scheme to support such research that cannot be (fully) funded out of existing research streams such as Research Councils.

Research into virtually all aspects of contemporary Gaelic linguistics (perhaps bar surface phonology and historical linguistics) is not well developed. Apart from specific research for SALT projects, research into fundamentals should be a priority.

Key topics for research include:

- Phonetic & Phonology: intonation, stress, prosody
- Morphology: derivation, compounding, contemporary case marking, long noun phrases
- Semantics: any
- Syntax: any
- Sociolinguistics: language change, use and acquisition and other topics relevant to codification and SALT development

5.3.4 Information Management

The flow and exchange of information is crucial. There have to be channels of information that inform the Gaelic-speaking public of new and planned developments and meaningful ways in which the public can interact with the “movers and shakers”. Modern means of communication such as web fora, e-newsletters and printed newsletters should be used to this purpose.

The flow of Gaelic-related information towards the non-Gaelic speaking public also needs to be improved, ideally by setting up a Gaelic hotline that provides support to bodies, organisations and companies on issues such as small translations and proofing of signage. Existing schemes in other countries such as Freagra (see Appendix 2) in Ireland provide excellent models.

5.4 Stage 2 - The SALT Centre

The key development at stage two is that of a centre of excellence in SALT. The network diagram shows the key dependencies for development and places the majority of SALT efforts into Stage 2. However, the actual approach is not quite as staged and there are certain immediate functions that require such a centre to be set up early.

For example, the centre will have an early role, both in the adaptation/build of a corpus and the production of SALT tools such as a lexical database, a lemmatiser and a generator.

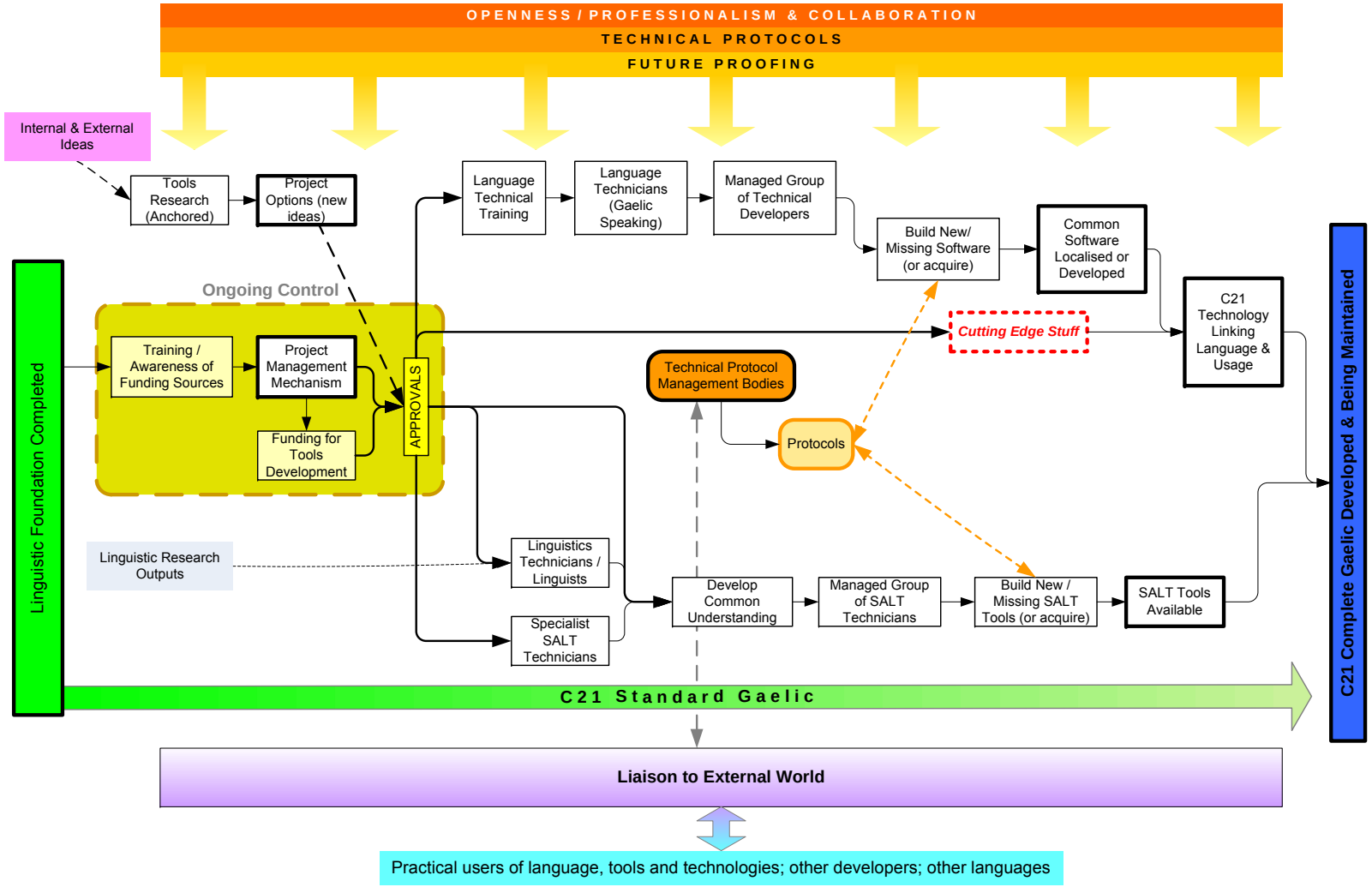


Figure 4 The SALT Centre and Downstream Development

There are a number of key documents produced in other countries that provide much technical detail that would be relevant to the Gaelic context as well. These should be carefully considered by such a future team, in particular (see Attachments for complete digital copies):

- Bilingual Software Standards & Guidelines (2006)
- Design Principles for the New Corpus for Ireland (2004)
- Information Technology and the Welsh Language: A Strategy Document (2006)
- Integrating NLP Tools for Basque in Text Editors
- SALTcymru (2008)

5.4.1 Setting up a Centre of Excellence

There are important considerations when setting up such a centre. Based on experience in other languages it is critical to establish some ground rules.

Within the management and control framework of the centre, the centre otherwise requires a great degree of operational freedom to respond quickly and flexibly to the ever-changing landscape of SALT.

To ensure continuity and sustainable development, the core functions of the SALT centre must be permanently funded. In practice this will require BnG to guarantee funding at certain levels for periods of more than a year (suggested timeframes are 5 year cycles).

Looking at the other funding needs of the SALT activities will be part of the remit of the core team. It is not unreasonable that they should be tasked with finding appropriate funding partners dependent on the outcomes of envisaged activity. However, there needs to be some underpinning of the research and development work. Experience in Wales shows that a lack of financial security can have a detrimental effect on the operations of such a centre. A model whereby in the longer term BnG acts as a guarantor for an otherwise self-funding SALT operation, would be a good solution.

In terms of skills, the core team must contain the following:

- Professional project management
- Fundraiser with experience in funding research[20]
- Lexicography
- (Technical) Gaelic translation
- SALT software development

Ideally all of these will be Gaelic-speaking. However, the overriding factor must be experience and technical skills. Access to training for the translators must be arranged early as there are currently no trained translators. Training should preferably be carried out by qualified and experienced Irish speakers due to the linguistic proximity of the two languages which result in highly similar problem areas. Also, to enable the long-term goal of fostering an indigenous skills base, there needs to be an early focus on developing access to training, either in language skills for non-speakers or in technical skills for speakers. A student placement scheme based at the proposed centre should also be considered.

[20] If the centre is associated with a university, it may be possible to utilise internal fundraising mechanisms.

Due to the nature of the work carried out by such a centre, it requires:

i. Close proximity to an existing established major university or universities with a wide academic offering. This will enable a maximum of intra- and inter-university collaboration and will facilitate access to resources.

ii. Proximity to a sizeable concentration of native Gaelic speakers (irrespective of training). With training, these may provide some of the future native skills base and act as valuable informants in Gaelic projects.

iii. Access to travel connections to engage in national or international events relevant to their work.

This centre should aim to establish itself as the focus for maintaining and further developing existing Gaelic SALT such as the Dearbhair or software translation projects. Taking ownership or at least control of the currently disparate projects will ensure better coordination (including terminological issues) and less duplication, foster the indigenous skills base and create a one-stop-shop for users looking for SALT.

Once the initial raft of tools has been taken on board and developed, the centre will utilise internal and external ideas that are validated through research to generate new projects. Early user testing, especially in case of completely new and untested ideas, will be an integral part. Constant reference to and compliance with international standards and protocols will be paramount. Approval for new projects will be via project management. The main emphasis should always be on tools that will be of practical benefit to as many users as possible.

The centre will also be in charge of promoting and disseminating developed tools[21]. Much greater use should be made of the web in this respect, especially in relation to improving familiarity with tools and their features. Video tutorials are both increasingly common and effective.

5.4.2 Typographical Tools

As modern Gaelic uses the core Latin alphabet, typography in the sense of font-related issues is not a common problem. The only exception is the use of artistic fonts unsuitable for Gaelic, which may warrant some general guidance to public sector bodies involved with Gaelic signage.

However, typography in the sense of typing Gaelic is a pressing need. Simple solutions such as the use of Irish keyboard layouts must be explained properly and promoted heavily.[22] Solutions to common typographical problems in non-Gaelic Office applications (such as preventing forced lower-case letters, e.g. in surnames such as MacGriogair or MacDòmhnaill) must also be addressed. Development of an auto-installer for people with limited computer literacy would be preferable. Due to the costs involved, promoting Gaelic hardware keyboards may prove difficult but the development of a specific Gaelic keyboard layout may be useful at a future point.

Given the fact that texting is so common-place amongst young people, the development of predictive Gaelic texting that can be installed across phones must be a priority. As there are currently limiting effects of using accented characters in text messages (this restricts them to 70 characters instead of 160), such software should have the option of stripping out accents if a message exceeds 70 characters to prevent the tool from being more limiting than the equivalent English version.

[21] This may, of course, be in conjunction with other Scottish Gaelic organisations.

[22] For an interactive display see www.microsoft.com/resources/msdn/goglobal/keyboards/kbdir.htm

Given that dictionary files for Gaelic spellcheckers and word prediction already exist (An Dearbhair, Penfriend), it should be relatively easy to adapt the technology to suit Gaelic. Given the domain and target audience, care should be taken to include words and abbreviations likely to be used by young people in preference to formal Gaelic. A competition in GME to suggest such could be used both to inform the database and to promote the future product.

5.4.3 Common Use Tools

Overall, the development of common use tools must focus much more on the use of Open Source software. Utilising Open Source software has a raft of benefits for the language:

- Opens up the pool of applications that can be utilised because Open Source software can be freely localised and does not require complex negotiations with owners of proprietary software. This pool of open software is maintained and added to by the global community and is increasingly widely-used. It allows lesser-used languages to provide a range of tools from Office software, web-browsers and email programs through to more sophisticated tools such as desktop publishing or operating systems.
- Allows the addition of language-specific tools and adaptations.
- Fosters the local skills pool.
- Adds financial benefits to the speaker community and increases the perceived economic value of the language.
- It either results in overall savings or in more resources being available for additional developments.
- Products are freely available to users who might be reluctant or unwilling to pay for proprietary software.

> Open Source software is increasingly widely used both privately, by business and even governments
>
> - Firefox has a 31%, Opera 5% market share in Europe[23]
> - OpenOffice has an estimated share of the commercial market of 15-20% and has been adopted by bodies such as the Singapore Ministry of Defence, the German Foreign Ministry and the French National Assembly.[24]

However, should opportunities arise for the localisation of proprietary tools at no or extremely little cost to the SALT centre, this should also be considered to further increase the amount of tools available to users.

5.4.3.1 Office Software and Operating Systems

There are currently two office software suites, Microsoft Office (completed but as yet unreleased) and OpenOffice and one operating system, Windows Vista.

The localisation of Microsoft products is both costly and usually associated with much delayed release dates for smaller languages. The Gaelic translations of Microsoft Vista and Office 2007 were initiated in 2007 but as yet not even the CLIP has been released (in contrast to, for example, the Welsh CLIP).[25] In addition, Microsoft has already released Vista's successor, Windows 7.

[23] Source: AT Internet www.atinternet.com

[24] Source: http://wiki.services.openoffice.org/wiki/Market_Share_Analysis

[25] Welsh CLIP http://office.microsoft.com/en-us/suites/HA103161311033.aspx

Future emphasis should be on developing Open Source software and ensuring cross-platform compatibility (Windows, MacOSX, Linux). A minority of users surveyed describe themselves as expert users of software. Particular emphasis should therefore be placed on commonly used tools such as Office applications, browsers, email programs and media software.[26] Initially less emphasis should be placed on localising operating systems as, according to the survey, a lack of computer literacy at the level of the Operating System is likely to make users reluctant to use a Gaelic OS. There are also issues surrounding technical support. Until such support becomes available through the medium of Gaelic, having to deal with Gaelic OS issues and English medium technical support may not be a popular option.

Use of Open Software should also facilitate dissemination as it would be freely available. Users are likely to be reluctant to commit themselves to purchasing expensive proprietary software.

Focus on commonly used Open Source tools:

- Office applications
- Web-browsers
- Email programs
- Media software

The levels of usage of localised Gaelic software should periodically be evaluated to identify gaps in dissemination or promotion. Numbers of downloads can be used as a simple, early indication of likely uptake. Should a piece of software become obsolete, support should not be continued without good reasons.

5.4.3.2 Proofing Tools

Control or agreement should be sought with the Dearbhair team to develop functionality, integration and compatibility of the software. Current problems with installation and operation of the Dearbhair must be addressed, in particular the problems the software has with learning new words.

Once lexical databases become available, expansion into Open Source software spell-checkers such as HunSpell should be tackled. If the Dearbhair project should prove to be unsupportive, an Open Source solution should be adopted immediately.

Feedback from and testing by users as regards functionality must be sought and carried out.

Once codification has completed, the data of existing spell-checkers will almost certainly need to be updated. It may be desirable to offer flexibility within spell-checkers in terms of the orthographies on offer as current digitisation projects frequently deal with variant orthographies.

[26] Exceptions might be made for software tools that would facilitiate the use of Scottish Gaelic in businesses as this is an extremely underdeveloped area of Scottish Gaelic in general.

Spell-Checkers

We have had a look at the three existing Gaelic spell-checkers and propose a simple method of assessing how useful they are. This is a developed formula where:

n = number of word-forms in the spellchecker.[27]

f = factor of words **:** forms specific to the language (we have used a value of 1:10 based on Irish spell-checkers.)

L = number of words in the language (Dwelly lists 78,000 and that is a reasonable basis to work with.)

z = functionality as an operable system. A low percentage score for ones that crash, are not updated, do not learn as they go, have no cross-platform compatibility.

So effectiveness of a spellchecker is $\frac{n}{f} \div L(\%) \times z$

For one of the current spell-checkers we have the following values: n = 547,000 and z = 20%.

Therefore content effectiveness is $\frac{547{,}000}{10} \div 78{,}000 \times 20\% = 14\%$

Leaving a final value for this particular tool of **14%**.

By comparison the same score for UK English in MS Word is well in excess of 90%. Even there it is not perfect because it is easy to teach the software **wrong** answers and perpetuate mistakes.

What does this mean? Well in practice the spell-checker used in the calculation is only 14% effective because of its limited functionalities and the size of the lexicon. The impact on the user community may be seen as being significant:

- Higher rework costs in possibly 86% of instances
- Poor quality of published materials
- Increased error rate, potentially leading to commercial consequences

However, in order to move such a tool into the realms of being satisfactory requires:

- Improved technology to make it more usable/reliable
- Bigger lexicon and more forms

[27] For example, *beinn* is a head-word; the related word-forms are *bheinn, beinne, bheinne, beanntan, bheanntan*, etc.

Overall:
The cost of not having a spellchecker can be seen as the **inverse** function of its rework. For the previously assessed Public Sector translation (see 5.3.1) we have a range of values representing 5% rework (if we achieve MS Word English Standard) to 86% (for the worst case we evaluated).

That means the spellchecker has an opportunity cost accordingly:

Capitalised that works out as an opportunity cost of **£23.6m** in the wider community.

Note: *The low estimates in the components assume that there are few problems and the population really doesn't need a spell-checker. On the other hand the high end emphasises the otherwise hidden costs throughout translated and other written texts that are reliant on poor standards.*

Optical Character Recognition (OCR) often caters for a wide range of languages as, at the simplest level, it only requires a definition of the character-set of a language and does not technically require localisation. However, integration of a spell-checker and/or lexical database into such software can improve the quality of the output. Again, as OCR will frequently deal with pre-standard systems, the option of using non-standard forms in such projects would be beneficial.

Best practice in grammar-checking currently requires large, tagged, linguistic corpora which are used in conjunction with the more traditional rules frameworks.

The development of such a tool fits within the wider strategy. However, Gaelic currently lacks a high quality corpus. Although there are currently no lexical databases for Gaelic either, there may be opportunities to begin developing a limited rules-based tool, perhaps with a view to a future hybrid system. Given the prior experience in building a rules-based grammar-checker for a Celtic language, a future Gaelic SALT centre should investigate collaborating with the Canolfan Bedwyr and Professor Scannell. Professor Scannell has built an Irish grammar-checker that is now available as a commercial product and has experimented with a similar Gaelic tool. For comparison, the Irish tool contains c.2000 rules, the Gaelic prototype c.200. Such a tool should also consider the question of high/low register in Gaelic to prevent alienating users that utilise the tool for informal written material (e.g. *dath na cloiche mòire* vs *dath na cloich(e) mhòr*).

Other more sophisticated tools are currently not feasible but should be considered by the centre as and when feasibility improves.

5.4.3.3 Terminology Tools

In conjunction with the terminology standardisation team, the centre must take on the creation of a national database of terminology. Collaboration with the Irish project Focal (see Appendix 2) should be seriously considered. However, due to popular demand, the scope of an equivalent Gaelic project should aim from the outset to include grammatical information, examples of usage, sound and pronunciation,[28] common terminology (with indications of regional variation) with the aim of a one-stop-shop for terminology. This will also counter the phenomenon that users will use such a resource to extract common terminology as much, if not more so, than technical terminology. This is currently an issue with Focal which does not (yet) provide sufficient context and examples of usage.

Indications of level of usage would be helpful, including for existing common terminology as users are currently left in the dark as to which items of terminology listed in dictionaries are commonly known where and which are not.

The output should be primarily digital, online and as a downloadable resource. Output should also include the data on place-names, surnames, names of bodies, etc to prevent (especially professional) users having to consult multiple sources simultaneously. Links should be put in place to Irish terminology sites, including Logainm (see Appendix 2).

In the short term, an effective way of collecting and disseminating terminology used and created by the Gaelic departments of the BBC should be found quickly. This should be publicly accessible via the internet to allow both BBC internal staff to "sing off the same sheet" and to support dissemination of terminology amongst the wider population. It should also be suggested to BBC Alba and MG ALBA that the contracts to independent production companies should include a clause that technical terminology used by the BBC (that has been made publicly available) should be adhered to as much as possible and that other relevant technical terminology used by the production companies should be supplied in the format of digital spreadsheets.

In view of this, co-operation with the Am Faclair Beag (AFB) project should also be considered. This project has an existing framework that may fulfil the majority of needs for such a resources.

Integration of dictionaries and thesauri into Office applications should also occur when and as these become available.

5.4.3.4 Computer-assisted Translation (CAT) Tools

Due to the various limiting factors in Gaelic translation (part-time, freelance, limited amount of work, lack of usage of existing CAT tools), the centre should select an appropriate Open Source CAT tool. They should then promote its use amongst Gaelic translators and agencies dealing with Gaelic translation. Adopting proprietary software such as SDL Trados, Lingo or MemoQ is not advisable. Although these are commonly found in the translation sector in mainstream languages, questions of cost and adaptability within a lesser-used language such as Gaelic make them a problematic choice.[29]

[28] Development of such a database that includes clear (IPA) phonetic transcriptions will also support the future development of speech technology which often requires phonetic lexica.

[29] Translation memory products for freelancers are in the region of €595-1840 (Trados), €620 (MemoQ) and US$599-1299 (Lingo).

In view of these considerations, in particular regarding future usefulness to translators outside UHI, the recent decision by UHI to invite tenders from providers of proprietary CAT software should be revisited urgently. A better solution could be to engage with Foras na Gaeilge and commission a joint ITT to adapt an open-source CAT tool in order to deliver improved features for Goidelic languages. Traslán, in conjunction with Foras na Gaeilge, is currently working on making translation memory files in the common TMX format available to Irish translators. Traslán encourages translators to try free tools such as OmegaT to familiarise themselves with the technology.

A collaborative project involving a free tool would have several advantages. A CAT tool for Gaelic would be acquired that would be affordable by all translators and a collaborative platform within the Goidelic community could be built, thereby enhancing resilience.

Engagement should also be sought with Scottish public sector organisations regularly involved in commissioning Gaelic translations (perhaps even into other languages)[30] to ensure that this translation tool (once developed) forms part of any issued requirement. This will have the following effects:

i. It will improve consistency between translation jobs and between translators.

ii. It will drive down costs to the Public Purse by deduplicating translation work.

Care must be take to explain clearly the proper use, advantages and pitfalls of CAT to manage expectations.

In collaboration with translators and public sector bodies (by whom the majority of translation work is commissioned) TMs should be collated, checked and made available free of charge for translators to assist CAT and to further promote the technology. The framework can be put in place at an early stage but implementation of the TM will have to await codification work being completed. It will also require a certain amount of updating as terminology standardisation progresses.

Gaelic Machine Translation (MT) is not feasible currently due to the lack of research and resources. In general expectations of MT are too high. MT between closely related languages may achieve accuracy levels of up to 95% but for distant languages (such as English to Gaelic) they are generally taken to be c.65% at best. It also remains most useful in extremely larger, technical translation projects which currently are rare in Gaelic. As the availability of data and basic resources progresses and if there is demand for the output that can be reasonably expected, the idea of joining on Open Source MT can be re-visited, in particular an Irish/Scottish Gaelic project due to the linguistic proximity.

Should such a project be considered at a future date, the EGGE/Numeral proposal (see page 77) could be re-evaluated.

5.4.4 Community Translation Projects

There is also an increasing pool of software, some proprietary, that provides community translation projects. These include tools like Google, Facebook (currently no Gaelic project). Where a case can be made for usefulness of a tool to the wider community, the centre should ideally take a controlling or leading role in the translation process to improve quality, standardise terminology and provide a wider range of tools.

[30] Essentially translation memory software is not language specific but matches strings of words/characters. Hence this could be equally applicable to Urdu, Polish etc.

Where Gaelic is currently not listed as a community translation project language, the centre should make an effort to have Gaelic included. The centre should also stress the importance of having equal levels of functionality in such projects, in particular the Google in Your Language project where currently the Gaelic project offers more limited functions than mainstream language interfaces.

Such sites can also be used as "dispersal engines" for new technical terminology.

5.4.4.1 Wikipedia

Despite controversies around accuracy issues in Wikipedia, the number of users consulting Wikipedia continues to grow, as does the number of Wikipedia projects. A 2007 report into Wikipedia usage amongst US citizens found that some 36% of Americans used Wikipedia as a source of information, with usage particularly high amongst young people (44% amongst 18-29) and college graduates (50%).

Usage amongst the smaller Wikis is likely to be less as people frequently expect that a corresponding article will contain less information. For example, the English article on Wales has 10,810 words; the German version has 2,150; the Irish 291 and the Gaelic a mere 99.

However, given the demographic limitations of the Gaelic speaker base, it is unlikely that there will be a Gaelic (up-to-date) encyclopaedia in the near future.

The following should therefore be considered:

- Set up a small team of dedicated editors to expand the Gaelic Wiki
- Develop a tag that indicates to users that the article is watched by an editorial team
- Enlist the help of Gaelic secondary schools and Gaelic HEI's to produce good Gaelic Wiki articles on a diverse range of topics
- Encourage participation from the wider community by providing basic information on how to participate responsibly

Except for the editorial costs, this could become a virtually universally accessible and up-to-date Gaelic encyclopaedia, with all the entailed benefits of such a tool, at a fraction of the cost of a printed volume.

5.4.5 Other Technologies

The centre should keep an open mind about other types of technologies and software. Games (including Open Source games), add-ons, plug-ins and other such applications are likely to be of particular interest. Mobile technologies are widely seen as a key technology area and therefore a close eye should be kept on this area to spot early possibilities for development.

One suggested project is the transformation of an existing digital dictionary such as Dwelly-d into a thesaurus. Such a project would require a corpus or lexical database to address ambiguity issues. However, it may be possible to develop a basic tool without a lexical database initially. Both Napier University and Professor Scannell (St. Louis) have indicated potential interest in such a project.

It will also be up to the centre to evaluate some of the less central suggestions produced in the creative workshops such as cross-university Gaelic support classes or the use of mobile technology to feed Gaelic-related information to users (see Section 4 and following).

5.4.6 Speech Technology

Text To Speech (TTS) technology is a form of speech synthesis that is not only of interest for mainstream languages but also, or perhaps especially, for minoritised languages. High-end TTS technology has numerous potential applications. Given the demographics of Gaelic and the learner community in Scotland, general pronunciation and reading problems, speech technology has much potential:

- As an educational tool. This includes adult education where learners frequently do not have access to good models of pronunciation.
- As a tool to widen access to written material by illiterate or semi-literate speakers and to improve literacy.
- As a support tool for children in GME where the home language is not Gaelic.
- As assistive technology for disabled users.

There are also possible applications outside the immediately obvious sectors. For example, in the tourism industry; or anywhere else where non-Gaelic speakers come into contact with the written word.

Along with the facility of modern speech synthesis to acquire regional accents, in theory this technology could also be used as an educational tool in the promotion of less-commonly represented or endangered Gaelic dialects.

Taking into account the potential impact, the centre should therefore embark on a Gaelic TTS project as soon as possible. It is important to bear in mind the fact that non-Gaelic speakers are likely to perceive such a Voice as a model. Also, previous experience with (English) speech technology has to date been largely negative in the Scottish context. It is therefore crucial that any release to the public is of high quality.

Although some pilot studies have taken place in the past, there is currently nothing in Gaelic that can be considered functional or industry standard. There are two fundamental approaches that can be taken in the case of Gaelic: a commercial or academic partner. Two groups have indicated interest in developing Gaelic TTS, CereProc (an Edinburgh-based company) and the Abair project (at Trinity College Dublin).

Either approach has both advantages and disadvantages. The main advantage with CereProc would be that the company is well-known for producing high-quality Voices and for being leading edge in terms of speech technology, including sophisticated aspects such as the integration of emotion. For historical reasons, the project at TCD is lagging slightly behind but is equally keen on developing cutting-edge technology, including the integration of emotion.

There are several main reasons why the main partner should be TCD. The TCD team has always been extremely keen to develop a Gaelic Voice and to develop an indigenous skills base. The linguistic proximity also means that much of the foundation work done in producing the Irish voices could be used in a Gaelic project. With a view to long-term development, including the development of more than one Voice, cooperation with TCD is also likely to be more sustainable.

The recommendation is therefore to join the Irish TTS project at an early date. However, CereProc as a possible third partner should also be investigated.

The reverse type of technology, Speech To Text (STT) also has considerable potential benefits. However, it is by far the more sophisticated of the two speech technologies and, also, due to the lack of existing foundations, not feasible at this stage. Therefore it should not be tackled as a priority. As more resources become available overall and as the technology advances, this may be revisited by the centre at a future point.

The Opportunity Cost/Value of Speech Synthesis

This is going to be usable by a whole raft of areas: assisting the disabled; as a teaching tool; providing speaking models; and telecoms automation.

Ignoring most of these, If we simply assume that some users need some form of Gaelic voice to support their activity (whether that is learning or to correct other material) we might suppose that having the technology available is a ready stand-in for a resource that is currently difficult to deliver 24/7 across a diverse geography.

Having such access would also be likely to increase the usage and demand as people got used to having it at their fingertips - even though today nobody is demanding it because Gaelic Voice Coaches are simply too thinly spread. This gives an opportunity cost/value as follows:

- Number of users In range 5,000 - 10,000
- Cost of 1 hour of Gaelic Voice coaching is £20-30
- Hours required per week is in the range 0.5 - 2
- Number of user weeks in year is 40

We can therefore derive an **NPV of £130m**.

5.4.7 Becoming Cutting Edge

One of the outcomes of the creative workshops was the desire for Gaelic to become leading-edge in some aspect to prevent the eternal pattern of trying to catch up with mainstream languages.

There are undoubtedly many different approaches that could be taken and it is not within the remit of this report to generate and evaluate all potential projects. However, one particular idea that might be of particular significance to Gaelic and other lesser-resourced languages is that of gaming. Such a project would have the added benefit of targeting language use especially amongst young people.

Gaming has been almost completely ignored by lesser-resourced languages in spite of its potential. Virtually all mass-market video/console games released today are localised in only a small handful of mainstream European and Asian languages and virtually never in languages with less than 50 million speakers. World-wide, of c.6,000 languages in total, less than 30 languages fall into the category of having that many speakers.

In the EU alone there are more than 45 million speakers of more than 30 minoritised languages, plus a number of relatively small current or future member languages like Maltese, Estonian or Icelandic.

The development costs of console games today range between £12m and £30m. Set against the increasing time spent by young people playing virtual games, there may be an opportunity here to set up a collaborative project to develop cutting-edge video games that would only be released in smaller/lesser-resourced languages. As there are currently almost no universities globally that offer specific degrees in games design and programming, this could be an opportunity for Scotland to take the lead in a completely new initiative if such an (English medium) degree were to specifically target the development of games in smaller languages.

6 Conclusion

It is clear from the foregoing that Gaelic needs to revisit the structures in place to control and manage the technical aspects of corpus development and SALT. This in turn will lead to developing a completely new set of arrangements for moving forward.

Current developments need to be critically reviewed and in many cases suspended or cancelled altogether. Failures of Quality Assurance to date have wasted significant amounts of resource that could have been better utilised. Overall this highlights the need for a strategic/planned approach, for language professionals to be put in charge of corpus development and SALT; and to move away the carrying out of corpus development and SALT from the schools' sector.

- Core codification of the language as soon as is practically possible
- Transferring ownership of the orthography to the beginnings of a Gaelic Academy, making the latter responsible for the Quality Assurance of the codification and terminology development
- Development of a gold standard corpus, managed by a Gaelic/Celtic HEI. (This alone will have a capital value to the Gaelic Language community in excess of £173m and can be delivered for a tiny fraction of that cost.)
- Setting up a Governance Framework that has adequate guaranteed funding to enable the baselining of the Language and Tools
- Temporary suspension of the publication of authorised publications prescribing orthography/lexical/grammar usage (e.g. An Seotal, Ainmean-àite na h-Alba). This does not mean these project should cease all work, rather that they will be engaged in coordinating with the new Gaelic Academy training in best practice other foundation activity. Once the fundamentals (including core codification) are in place, these projects should then resume.

  If possible, delay the launch of new proofing tools until codification is complete and can be implemented.
- Beyond that adherence to International Protocols and Standards must be achieved at all levels.

In addition to these core recommendations, there are various additional recommendations, suggestions and ideas which can be found in the relevant sections of this report.

Collaboration with the following projects and partners should also commence at the earliest opportunity:

- Canolfan Bedwyr (Welsh SALT centre)
- Fiontar (Irish SALT centre)
- Foras na Gaeilge (Irish cross-border development agency)
- NCI (New Corpus for Ireland project)
- Professor Scannell (Professor of Computing, University of St. Louis, Missouri)
- Traslán (Translation company working in conjunction with Foras na Gaeilge)

Training and development of core teams must take place to engage in the technical work of Terminology Standardisation, and Corpus work

Finally wider aspects such as engaging with younger people to develop appropriate supportive technologies in areas like games and texting should be encouraged. All of this downstream activity should be professionally project-managed to ensure that there is cohesion and optimisation of resources and sharing of the collective output.

This represents a major challenge but if the proposals are followed Gaelic will take its place among the leading minority languages, rather than fire-fighting. The outcome can do nothing but good for the wider development and uptake of the language in the 21st century.

Appendix 1

Index of Promised Deliverables

Deliverable	Commentary	Location
A comprehensive list of existing tools, technologies and services.	Tables of projects in alphabetical order with a topical index.	Appendix 2
An assessment of the value/ effectiveness of each of the above.	Various assessments based on: ▪ Survey results ▪ Workshop output ▪ Economic appraisal of certain components (especially hidden impacts on Gaelic users) ▪ Comparison with similar items in other countries ▪ 3rd party research and reports This has been dealt with by combination of: ▪ Narrative according to topic ▪ Commentaries (yellow text boxes) ▪ Survey and workshop reports	 Throughout Throughout Appendix 3 Appendix 4
A consolidated view of current and proposed developments, validated as to their potential attributes and delivery, together with an assessment of the likely value to Gaelic Corpus Planning.	There is no section to deal specifically with a consolidated view. Instead, this has been absorbed into the Roadmap for Gaelic. The likely value is highlighted by reference to the interdependencies shown in Figure 2, Figure 3, and Figure 4. Current and proposed developments have been dealt with by a combination of: ▪ Narrative according to topic ▪ Commentaries (yellow text boxes)	Section 5 Sections 5.1; 5.3; and 5.4 Throughout Throughout
A Consolidated Needs Analysis looking to the future from BnG, other stakeholders and including alternatives and 'new' ideas generated by our work.	This is the main thrust of the Roadmap for Gaelic, achieved by comparing the General Schema with ▪ Current Gaelic World ▪ Aspirations of Gaelic Users ▪ Survey responses ▪ Workshop output It not only includes projects and tools but also the frameworks and governance that will facilitate the overall development.	Section 5 Section 2 Section 3 Section 4 Appendix 3 Appendix 4
A Gap Analysis highlighting the way(s) forward with emphasis on the value chain.	Combined into the Roadmap, see above.	

Appendix 2

List of Projects

This appendix lists all projects that have been investigated for this report.

- Table I is an index of all projects (with short descriptions), sorted according to type.
- Table II (alphabetically) lists, the currently existing and ongoing Gaelic/Scottish projects
- Table III (alphabetically) lists currently planned Gaelic projects
- Table IV (alphabetically) lists all projects that were investigated in Wales, Ireland, Northern Ireland and the Basque Country

Table I - Index of All Projects

Type & Name	Short description	Page
Corpus/digital archives		
CEG	Cronfa Electrone.g. o Gymraeg, a Welsh (tagged) corpus	105
CELT	Corpus of Electronic Texts, digital archive of Old, Middle and Modern Irish texts	75
Corpas na Gaeilge	Corpus (untagged) of Early Modern and Modern Irish	108
Corpas na Gàidhlig	Corpus (planned) of Gaelic, see DASG	75
eDIL	Electronic Dictionary of the Irish Language, corpus of Old and Middle Irish	110
DASG	Digital Archive of Scottish Gaelic, digitisation project	75
LER-BIML	Gaelic corpus (untagged)	79
Nancy Dorian	“Corpus” of East Sutherland and Black Isle Gaelic	81
NCI	New Corpus (tagged) for Ireland (Irish and Hiberno English)	128
NLS	National Library of Scotland, Gaelic digitisation project	82
SCOTS	Scots corpus project (untagged)	87
Tobar na Gaedhilge	Corpus (untagged) of Modern Irish and Gaelic	135
Will Lamb	Gaelic (tagged) corpus	93
Dictionaries		
Acmhainn	Irish online termbase	98
AFB	Am Faclair Beag, online dictionary	72
Collins	Modern Irish dictionaries	107
Dwelly-d	Dwelly Digiteach, online version of Dwelly’s Dictionary	77
Elhuyar	Publisher of Basque scientific literature and dictionaries	110
Euskalterm	Online database of technical Basque	134

Type & Name	Short description	Page
Faclair Bun-tùsach	(Proposed) bidirectional Gaelic dictionary	94
Faclair na Gàidhlig	Historical Dictionary of Gaelic	77
Faclair na Pàrlamaid	Dictionary of parliamentary and governmental terminology	78
Faclair nan Gnàthasan-cainnte	Database of Gaelic idioms	78
Focal	Online database of Irish terminology	118
Grammar dictionary	(Proposed) Gaelic dictionary of grammar and grammatical terms	95
HDSG	Historical Dictionary of Scottish Gaelic, see DASG	75
IATE	European online termbase for EU working languages	118
NEID	New English Irish dictionary	128
Pròiseact Comhairle	Dictionary of terminology for local government	86
Stòr-dàta	Gaelic online terminology database	88
Talking Dictionary	(Proposed) Gaelic talking dictionary	94
T-Rex	Gaelic thesaurus	94
Groups/centres		
Canolfan Bedwyr	Welsh speech and language technology centre	102
Elhuyar	Research & development of Basque SALT, linguistic consultancy	110
Napier University	Development of Gaelic educational resources and Gaelic SALT	81
SALTCymru	Welsh SALT networking centre	132
TELI	The European Language Initiative (Gaelic and other languages)	88
Traslán	Irish translation agency with involvement in CAT research and development	135
University of Abertay[31]	Digital heritage projects	91
University of Dundee	Assistive technologies	91
University of Edinburgh	Centre for Speech Technology Research	92
University of St Andrews	Corpus technology	92
Information management		
LinkLine	Welsh information hotline	127
Freagra	Irish information hotline	122
Research & development		
ASGP	Arizona Gaelic Syntax Project, linguistic research	74
Euskara Institutua	Basque centre for linguistic research	115

[31] The university projects and departments listed here are those that do not specifically deal with Gaelic.

Type & Name	Short description	Page
Fiontar	Irish centre for IT and management	117
Ixa	Basque research and development group	124
Technology - Proofing		
An Dearbhair	Gaelic spell-checker (Windows)	73
An Gramadóir	Irish grammar-checker	132
Cysgliad	Welsh spell- and grammar-checker	109
GaelSpell	Irish spell-checker	132
GaidhealSpell	Gaelic spell-checker (MacOSX)	78
Grammar-checker	(Proposed) Gaelic grammar-checker	94
Roy Wentworth's spell-checker	Gaelic spell-checking workaround	86
Technology - Software		
GiyL	Google in your Language, interface translation project	79
Microsoft Windows & Office	Microsoft Windows and Office localisation project	80
OpenOffice	OpenOffice localisation project	83
Open Source	Open Source software projects (Welsh)	129
Opera	Opera (web-browser) localisation project	83
phpBB	Forum software localisation project (Gaelic)	85
Ubuntu	Open source operating system localisation (Gaelic and others)	91
Technology - Speech		
Abair	Irish speech synthesis	96
CereProc	(Potential) partner for Gaelic speech synthesis	94
TTS (J Berry)	Gaelic text to speech project	89
TTS (M Wolters)	Gaelic text to speech project	90
TTS (Murray)	Gaelic text to speech project	90
WISPR	Welsh speech synthesis project	138
Technology - Other		
EGGE & Numeral	Gaelic machine translation prototype with a basic normaliser for numbers	77
Less-Mess	Onscreen keyboard software	126
MacOSX keyboard	Gaelic keyboard layout for MacOSX	80
Maes-T	Welsh terminology development management software	127
Penfriend	Gaelic word prediction software	84
Pools-T	Software tools for language learning and teaching (Gaelic and others)	85

Table II - Existing Gaelic Projects

Name	Description
AÀA	Ainmean-Àite na h-Alba (AÀA) superseded the previous body, the Gaelic Names Liaison Committee (formed in 2000). Its main goal is to research Gaelic place-names required for bilingual signage and to publish their findings. The project is currently only funded until March 2011 and it is not known whether there will be funding after this date or whether the project will end. AÀA is funded by Bòrd na Gàidhlig with contributions from Highland Council and Argyll & Bute Council. Requests for research from other bodies are charged. The database is currently not available to the public and contains only about 100 entries as most of the work carried out to date has focussed on designing and testing the database, establishing working principles and carrying out the necessary research. To date, approximately 1600 place-names, the majority of which are located in the Highlands, have been researched. Most of these are in response to requests from the authorities who require place-names for signage purposes. The public output will be via the AÀA website, partially bilingual, and will include genitive forms (but not the gender of opaque place-names) and general guidance on where to obtain advice on place-names. Documentation regarding orthographic principles adopted and promulgated by AÀA relating to place-names and street-names is also publicly available on the website. At the moment, the place-name information researched by AÀA is available in PDF format spread across a number of files. A number of typographical errors[32] occur in some of the publicly available files but apparently these errors are unlikely to be present in the underlying database. Robust procedures for proofing should be put in place prior to publishing on order to avoid disseminating erroneous forms. AÀA, Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ Rosemary Gibson, rgibson@ainmean-aite.org; www.ainmean-aite.org

[32] For example, *Allt **A'Ch**ruinn, C**eo**l Reatha, **A'Mh**ormhaich* (in A87 Bilingual Signing Scheme, www.gaelicplacenames.org/UserFiles/File/A87 Bilingual Signing Scheme.pdf)

Name	Description
AFB	Am Faclair Beag (AFB) is the follow-on project of Dwelly-d and is also privately funded. At its core it contains the original Dwelly-d data but also has a new database for modern terminology and advanced features such as: ▪ A voting system for registered and vetted native speakers, enabling them to pass judgement on their knowldge of items in the dictionaries. The aim is to evaluate which Gaelic words are (still) in use and where, which would be difficult to do via a real-life survey. The voting system is fully implemented and currently 17 voting users (representing Lewis, Ross, Skye and Tiree Gaelic) have been recruited. ▪ A framework (implemented) for an online dictionary with extensive grammatical information, including phonetic transcriptions and sound. This framework is simultaneously helping build a lexical database by recording and labelling word forms associated with root words. ▪ Tools to support editors in generating content. This includes a word form generator that is capable of predicting word forms for regular verbs, regular adjectives and those noun forms that can be predicted by rules. ▪ The function that sorts results by relevance has also been upgraded and is now capable of being trained by users to provide better results. The project aims to bring together various terminology resources. To date (September 2009) it contains Dwelly-d and the 23,000 entries of the Faclair nan Gnàthasan-cainnte (see page 78). Currently there are only two editors adding new content when time permits. This project is funded privately and by donations. Akerbeltz, 1/2 47 Wilton Street, Glaschu G20 6RT Michael Bauer, fios@akerbeltz.org; William Robertson, w.robertson@cairnwater.co.uk; www.faclair.com

Name	Description
An Dearbhair	A Gaelic spellchecker produced by TELI for LTS. The first version was launched in 2006, available free of charge for download. Suitable for Windows & Microsoft Office only. GOC orthography. No precise headword count (compiled as a text file), c.550,000 word-forms[33]. The latest released version is compatible with Vista but LTS notes that installing An Dearbhair is not compatible with also having Roy Wentworth's spellchecker installed. As of May 2009, work has been completed to integrate An Dearbhair into Office and to produce MacOSX and OpenOffice versions. Release date unknown. LTS, 58 Sràid MhicDhonnchaidh, Glaschu G2 8DU Annie NicNèill, a.macneil@ltscotland.org.uk; www.ltscotland.org.uk/gaidhlig/taic/goireaseile/gaelspell.asp
An Seotal	An Seotal was set up by Stòrlann in August 2007 to standardise Gaelic terminology. Originally set to run until 2008, the project has been extended to 2011. The online database currently contains approximately 500 terms, with emphasis of the project currently on scientific and mathematical terminology. An Seotal also deals with requests from teachers for specific items of terminology. Existing and conflicting terminology is collected and passed to a panel of 2-3 volunteer teachers for comments and suggestions. If no consensus emerges at this stage, it passed to the Advisory Panel for a final decision. The Advisory Panel of 8 (3 are Stòrlann staff) meets once every 2 months (members of the Advisory Panel carry out this work part-time along their other work commitments). In the database, parts of speech are marked and grammatical information on inflections is given. The team has the use of an in-house translator and project officer but there appears to be little or no input from trained terminologists or lexicographers or experts outside the education sector An Seotal, 11/12 Acarsaid, Cidhe Sràid Chrombail, Steòrnabhagh, Eilean Leòdhais HS1 2DF Criosaidh NicRath, chrissiemacrae@storlann.co.uk; www.anseotal.org.uk

[33] Based on the correlation of the Irish spell-checker with 33,000 headwords and 320,000 forms, this is the equivalant of approximately 54,000 headwords.

Name	Description
ASGP	The University of Arizona has a sizeable Gaelic research project, the Arizona Scottish Gaelic Syntax Project (ASGP). Its main focus is the study of Gaelic grammar, in particular its syntax. The project is headed by Prof Andrew Carnie, whose background is in Irish but who has a strong interest in Gaelic. The syntax research was funded by the National Science Foundation to critically evaluate existing research into Gaelic syntax and produce a database of web-compatible and searchable tokens. So far over 200 hours of native speech have been recorded and part processed. Not yet completed are the interlinear versions, the related database, online publication, etc. The project is due to be completed in 2010. The centre has also applied for a sizeable grant to further its research into Gaelic phonology and sound production. Researchers are also volunteering time to produce a comprehensive online grammar of Gaelic. This project is currently in its initial phase. University of Arizona, Dept. of Linguistics, Douglass Bldg, Room 200E, Tucson. Arizona 85721, USA Prof. Andrew Carnie, carnie@u.arizona.edu; www.dingo.sbs.arizona.edu/~Gaelic
BBC & MG ALBA	The various bodies (BBC Alba, BBC Gàidhlig, MG ALBA, Radio nan Gàidheal) involved in producing and commissioning Gaelic content for broadcasting are not themselves involved in the development of SALT. However, as BBC Alba and BBC Radio nan Gàidheal constitute the vast majority of available Gaelic media, they play a vital role in creating and disseminating terminology. Most terminology is determined by individual editors either in isolation or through consultation with on- or off-site colleagues. A small part of this terminology is held in an online wordlist called Facail Fheumail[34] which is not maintained regularly. Other than that, no central collection and dissemination facilities, either within or outwith the various BBC departments exist. It currently (Oct 2009) contains just over 500 entries, ranging from technical terms and place-names to everyday terminology. It would appear that there are currently no contractual obligations on independent production companies regarding the use of "established BBC terminology" or the creation of terminology lists to be supplied with the programmes. BBC Alba, Pacific Quay, Glaschu G51 1DA Mairead Màiri Mhoireach, margaret.mary@bbc.co.uk

[34] http://www.bbc.co.uk/scotland/alba/naidheachdan/facail/?letter=a

Name	Description
CELT	The CELT Corpus is not technically a Gaelic corpus but rather a corpus of Old, Middle and Modern Irish texts containing some 12.5 million words. It does contain a small amount of early Gaelic material, however, such as the Gaelic Notes in the Book of Deer. As such it may be of interest to any future Gaelic corpus project aiming to cover a wider historical period. Coláiste na hOllscoile Corcaigh, Corcaigh, Ireland Beatrix Färber b.faerber@ucc.ie; http://celt.ucc.ie/index.html
DASG	The Digital Archive of Scottish Gaelic (DASG) project was set up in 2006 by the Department of Celtic (and Gaelic) at the University of Glasgow. Its main goals are to: ▪ preserve and enhance the archive of the Historical Dictionary of Scottish Gaelic (HDSG, see below) ▪ digitise and publish the HDSG materials ▪ contribute to the Faclair na Gàidhlig project (see page 77) The project is collaborating with the SCOTS project (q.v.) on various issues to do with historical dictionaries, digitisation and corpus building. **Corpas na Gàidhlig** DASG, which contributes to the Faclair na Gàidhlig (q.v.) project, is working towards developing a corpus covering the historical period (beginning with the Book of Deer) up until the 21st century based initially on c.220 texts. Work began in October 2008 to digitise the material; to date (June 2009) over 20 books (20th and 21st century prose) have been scanned and 5 publications (c.210,900 words) have been digitised and proofed. Work on the actual corpus engine for DASG and Faclair na Gàidhlig has not yet begun. However, a substantial part of the necessary funding has been earmarked in the Faclair na Gàidhlig funding for 3 FTE corpus assistants (to carry out digitisation and proofing) and an IT specialist to be employed in 2009. Current plans envisage building a completely new corpus engine. The project is in contact with the National Library of Scotland regarding their Gaelic digitisation project (and with other individuals and organisations) to explore possible opportunities for collaboration. Once the work contributing to Faclair na Gàidhlig has been completed, it is envisaged that DASG will: ▪ add a wider range and number of materials, including audio material ▪ develop a tagged corpus

Name	Description
	The Historical Dictionary of Scottish Gaelic (HDSG) The Department of Celtic at the University of Glasgow collected a substantial amount of material in fieldwork between 1966-1996. It contains: ▪ questionnaires ▪ word-lists ▪ recordings on reel-tapes and cassettes; including some transcriptions ▪ paper slips (c. 500,000) ▪ other materials These are from a wide range of dialects, including material from less well-researched areas such as Nova Scotia and Kintyre. They cover a wide range of traditional areas and topics such as animal husbandry, fishing, past-times, tools, etc. This material was to be included in the Historical Dictionary of Scottish Gaelic (HDSG). However, work on the dictionary was formally abandoned in 1996. The focus of work on the HDSG materials is currently on the questionnaires and wordlists, the vast majority of which have been manually retyped in Word. It is planned to enter these materials into an online database. It is envisaged that at least some of this material will feed into Faclair na Gàidhlig. Roinn na Ceiltise is na Gàidhlig, Oilthigh Ghlaschu, 3 Gàrraidhean an Oilthigh G12 8QQ An t-Oll. Rob Ó Maolalaigh rom@celtic.arts.gla.ac.uk; www.gla.ac.uk/departments/celtic/projects/digitalarchiveofscottishgaelicdasg

Name	Description
Dwelly-d	The aim of this project was to produce a digital version of Dwelly's Gaelic-English Dictionary that could be available for searches online. The digitisation was handled by Michael Bauer (starting in 1998), the data migration, cleansing, development of the database (MySQL) and online interface by William Robertson (from 2008 onwards). Permission was obtained and minor aspects of the dictionary were edited in the digitisation process, such as taxonomy and some irregularities. It went live in January 2009. The dictionary contains c.78,000 entries and can be searched in either direction. The interface offers a variety of search options. Although capable of handling graphics, the illustrations have not been added to date due to monetary constraints. The main outstanding issue is the lack of user-friendly help functions to support the available features. This project is funded privately and by donations. Akerbeltz, 1/2 47 Wilton Street, Glaschu G20 6RT Michael Bauer, fios@akerbeltz.org; William Robertson, w.robertson@cairnwater.co.uk; www.dwelly.info
EGGE & Numeral	John Bruce, a retired software developer, has been working on various software projects. Numeral normalises Gaelic numbers following the decimal system (e.g. 15 → còig deug). It currently can handle only numbers without nouns. It is currently being evaluated by the digital archive of Highlands and Islands' culture, Am Baile,[35] for possible use on its site. English to Gaelic, Gaelic to English (EGGE) is a limited-domain machine translation prototype. Development is not particularly well documented and appears to be using a purely rules based approach. John Bruce, 33 Kintail Place, Inbhir Pheofharain IV15 9RL; kintailatv@yahoo.co.uk
Faclair na Gàidhlig	Faclair na Gàidhlig (FnaG) is an inter-university project between Glasgow University, SMO, Strathclyde University, the University of Edinburgh and the University of Aberdeen. The ultimate goal is the production of a historical dictionary of Scottish Gaelic based on a full-text atabase of c.220 texts, but also including other sources, for example, from DASG (q.v.). The initial project aims to cover the entire historical period from the Book of Deer up to the 21st century. The first phase of the project was completed at the University of Edinburgh. Work has commenced at Glasgow University to digitise c.220 Gaelic texts and publications (see DASG). Faclair na Gàidhlig, Fàs, Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ Lorna Pike, mail@faclair.ac.uk; http://www.faclair.ac.uk/

[35] See www.ambaile.org.uk

Name	Description
Faclair na Pàrlamaid	A dictionary of governmental and parliamentary terminology containing c.6,000 headwords was published in 2001. It was produced by TELI and based on a previous dictionary produced for the Welsh Assembly. The initial version received financial support from Urras Brosnachaidh na Gàidhlig and Comunn Gàidhlig Inbhir Nis and was available free of charge. An online version was produced in 2002. A second online version (with an expanded database and an inquiry service) was launched in 2007, funded by Bòrd na Gàidhlig. There has been little further development since. TELI, PO Box 1901, Milton Keynes, MK19 6DN Leo McNeir, mcneir@waitrose.com, www.leomcneir.com; www.scotland.gov.uk/dictionary/_bin/
Faclair nan Gnàthasan-cainnte	The Faclair nan Gnàthasan-cainnte is an online Gaelic-English searchable database of c.20,000 Gaelic idioms and expressions. It is currently the only resource that focuses on idioms and expressions rather than pure terminology. The database can be searched in both directions. The data has recently been integrated into AFB. However, some data cleansing remains outstanding. Akerbeltz, 1/2 47 Wilton Street, Glaschu G20 6RT Michael Bauer, fios@akerbeltz.org; www.akerbeltz.org/faclair/rannsachadh.php
GaidhealSpell	A Scottish Gaelic spell-checker produced by Everson Teo. Corpus based lexicon. Available free of charge for download. Suitable for various applications under MacOSX only. Non-GOC orthography. No exact headword count (based on a 132,000 word corpus), approx. 47,000 word-forms. Everson Teoranta, Cnoc na Sceiche, Leac an Anfa, Cathair na Mart, Co. Mhaigh Eo, Ireland Michael Everson, everson@evertype.com; www.evertype.com/software/macgaidhealspell/

Name	Description
Google in Your Language	The Google in your Language project is a volunteer-based online translation project that was set up in 2001 to make the Google interface available in non-mainstream languages. Participation is free to anyone. Not all features of the Google package are available for translation in each language. The Gaelic Main Search Site interface was mostly translated by a single translator and maintained between 2003 and 2007. The Main Search Site and Help Pages were the only GiyL projects that were at some point fully translated although partial translations exist for some of the other projects. The Gaelic translation has since lain mostly dormant due to a lack of skilled volunteer translators and some problems with support from Google in dealing with rogue translators and maintenance. Current status of translation (June 2009) is Main Site Help Pages (79%), Main Search Site (65%), Groups UI (53%), Wireless (46%), Orkut Mobile (26%), User Distributed Search (1%), Google Map Maker, Knol, Orkut Frontend Templates, Picasa2, Wireless Transcoder (all 0%). Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA; www.google.com/transconsole
LER-BIML	Funded by an EPSRC research grant, the Language Engineering Resources for the British Indigenous Minority Languages (LER-BIML) Project set out to survey existing language-engineering tools and end-user needs, develop EAGLES-conformant tags for the Celtic languages (and Scots) and develop a small 80,000 word corpus for Gaelic and Welsh based on spoken data and tag one of them using the tags identified. The project had a budget of £61,100 and ran for 17 months between 2002 and 2003. It involved Sabhal Mòr Ostaig, Bangor University, QUB, University College Cork and the University of Bedfordshire. Points to note from the survey are a high percentage of respondents (over 86 out of 127 in an open web survey) were keen to see Gaelic corpus tools and that a bilingual corpus was favoured. As regards to the tags, it was found that the guidelines developed for the National Corpus of Irish were compatible for Gaelic, so no new tagset was developed. The finished Gaelic corpus contains only 5 transcribed texts (2 sermons, one lecture, one talk and one informal conversation), numbering approximately 23,000 in total. The smaller size was due to the transcriber being much slower than expected, leading to the budget being depleted much quicker. The Welsh, rather than the Gaelic corpus was tagged with help from the University of Bangor team and their existing corpus. The original audio material, including the untranscribed material, is held at SMO. The University of Lancaster has the University Centre for Computer Corpus Research on Language (UCREL) whose scope formally includes non-English and minority languages. Their wider experience may be of use for gaining a wider perspective prior to commencing development. The project ended in 2003 and there are no current plans to develop it further but Dr Andrew Wilson has indicated that there is interest in participating in any future corpus projects. Dept. of Linguistics and Modern English Language, Lancaster University, Lancaster LA1 4YT Dr Andrew Wilson, eiaaw@exchange.lancs.ac.uk; www.ling.lancs.ac.uk/biml

Name	Description
MacOSX Gaelic Keyboard	Professor Andrew Carnie has produced a simplified keyboard layout for users of Mac OSX (10.3 or later). In the re-designed form, vowels with a grave are accessible by a simpler, two-key combination (CTRL plus the vowel), removing the need to press the Option key on Mac keyboards. This can be downloaded free of charge but requires a small amount of manual configuration during installation. University of Arizona, Dept. of Linguistics, Douglass Bldg, Room 200E, Tucson. Arizona 85721, USA Prof. Andrew Carnie, carnie@u.arizona.edu; www.dingo.sbs.arizona.edu/~carnie/publications/GaelKeyboard.html
Microsoft Windows & Office	A project to localise Microsoft Windows Vista and Office 2007 was announced in February 2007 by Microsoft and Bòrd na Gàidhlig. Initially announced to involve LTS and Strathclyde University, the project effectively appears to have been run by TELI alone. A Community Language Interface Package (CLIP[36]) was produced in 2007 with some minor input from the community. The CLIP is currently not available from the Microsoft website. In 2008 TELI created a Gaelic User Interface Package for Microsoft. According to TELI (May 2009) the full translation of Vista has been completed and is with Microsoft. Release dates are not known. All Microsoft/TELI projects were funded by Bòrd na Gàidhlig and Microsoft. TELI, PO Box 1901, Milton Keynes, MK19 6DN Leo McNeir, mcneir@waitrose.com, www.leomcneir.com

[36] Microsoft CLIPs exist for various, smaller languages. When downloaded and installed, they render part of the Microsoft software interface in the respective language.

Name	Description
Nancy Dorian	Nancy Dorian is an American linguist at Bryn Mawr College, Pennsylvania, who researched Gaelic dialects on the Black Isle and East Sutherland from the 1960s onwards. Her Gaelic research culminated in a number of publications, including *East Sutherland Gaelic* published by the Dublin Institute for Advanced Studies in 1978. Much of the material she collected from various sources was transcribed into phonetic annotations and some tagged for various types of linguistic features. Copies of some of the audio material have been sent to the School of Scottish Studies and SMO over the years. There are some issues regarding copyright about some of the (unpublished) material. Although perhaps technically not a written corpus immediately usable, it provides a wealth of recorded material of some now moribund dialects that should find its way into a corpus of spoken Gaelic. She is unsure of what her personal input to developing a corpus might be but overall is keen on it being developed. Nancy Dorian, 1810 Harpswell Neck Rd., Harpswell, ME 04079, USA; ndorian@gwi.net
Napier University	Staff from the School of Computing are involved in a number of Gaelic-related projects. The main emphasis is on developing educational tools for GME but there is also interest other areas. Work is currently being done on a word generator to produce a tool suggesting appropriate grammatical forms and lemmatisers that can be linked to digital dictionaries. There is strong interest in limited domain speech synthesis. The department also has some interesting projects on semantic networks that could be of interest to Gaelic.[37] The department is a strong believer in sharing knowledge and resources with the aim to stimulate the production of additional tools. It is currently the only known university in Scotland that utilises non-Gaelic students for Gaelic-related project work. School of Computing, Napier University, Room C62, Colinton Road, Dùn Èideann EH10 5DT Alistair Lawson, al.lawson@napier.ac.uk; www.soc.napier.ac.uk

[37] A semantic network is a visual representation of related meanings words (and could be described as a “thesaurus on steroids”).

Name	Description
NLS	The National Library of Scotland has, as of September 2009, embarked on a project to digitise some 3,000 Gaelic out-of-copyright books. These will not be proofread but available online in a variety of formats (PDF, Read Online, Text). Several hundred of them are currently already available on the Internet Archive and will be available on the NLS website at a future point too. NLS has indicated that there is potentially room for co-operation with a Gaelic corpus project. Director of Collections and Research, NLS, George IV Bridge, Dùn Èideann EH1 1 EW Cate Newton, c.newton@nls.uk; www.archive.org
OpenOffice	OpenOffice (OO) is an open-source office applications package. Originally called StarOffice and developed by Sun Microsystems, the source code was released in 2000 with the stated aim to reduce the global dependence on Microsoft Office. The current version is 3.2, running on Windows, Mac OSX and Linux. It comprises a word processing suite, a spreadsheet application, a presentation package, a database package, graphics software and a tool for creating mathematical formulae. It is currently available in 80 languages, including Irish (v2.1), Breton (v2.4), Gaelic (v1.1) and Welsh (2.0). The original Gaelic version (1.1) was produced by LTS in 2004/05. Cànan was contracted by LTS early in 2007 to produce a Gaelic version of the (then) current 2.4 version of OpenOffice. The Gaelic translation project has been completed and handed to LTS for reviewing and is expected to be released some time in 2009. The release version will be 3.2. The aim is to integrate the Dearbhair into OO but Cànan has not received the required files to date (June 2009). Cànan employed a panel of teachers, Stòrlann and various experts to deal with terminology issues. Limited contact and concordance took place between the OO project and the Windows Vista project. Cànan is currently trying to broach the subject of a maintenance agreement to provide regular updates. Pootle, an Open Source CAT tool designed for software localisation, was used during the project and although there are currently no plans for the TM data it could be extracted and made available. Cànan, Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ Shirley Grant, shirley@canan.co.uk; www.canan.co.uk

Name	Description
OpenOffice	OpenOffice (OO) is an open-source office applications package. Originally called StarOffice and developed by Sun Microsystems, the source code was released in 2000 with the stated aim to reduce the global dependence on Microsoft Office. The current version is 3.2, running on Windows, Mac OSX and Linux. It comprises a word processing suite, a spreadsheet application, a presentation package, a database package, graphics software and a tool for creating mathematical formulae. It is currently available in 80 languages, including Irish (v2.1), Breton (v2.4), Gaelic (v1.1) and Welsh (2.0). The original Gaelic version (1.1) was produced by LTS in 2004/05. Cànan was contracted by LTS early in 2007 to produce a Gaelic version of the (then) current 2.4 version of OpenOffice. The Gaelic translation project has been completed and handed to LTS for reviewing and is expected to be released some time in 2009. The release version will be 3.2. The aim is to integrate the Dearbhair into OO but Cànan has not received the required files to date (June 2009). Cànan employed a panel of teachers, Stòrlann and various experts to deal with terminology issues. Limited contact and concordance took place between the OO project and the Windows Vista project. Cànan is currently trying to broach the subject of a maintenance agreement to provide regular updates. Pootle, an Open Source CAT tool designed for software localisation, was used during the project and although there are currently no plans for the TM data it could be extracted and made available. Cànan, Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ Shirley Grant, shirley@canan.co.uk; www.canan.co.uk
Opera	Opera is a free cross-platform web-browser that was first released in 1996. In 2000 Opera Software and DART[38] announced the release of Opera 4.02 for Windows in Gaelic, Irish, Welsh and Breton. The majority of the work for Gaelic was carried out by SMO staff in their personal time. The Welsh and Gaelic projects continued until version 6.05 in 2002. The project has been dormant since. Interest exists to revive the project but no steps have been taken to date due to a lack of time. It is estimated that updating the Gaelic version to the most recent release of Opera (10.X) would require approximately 2 months of full-time work. Opera Software ASA, Waldemar Thranes gate 98, 0175 Oslo, Norway Gaelic add-on at http://arc.opera.com/pub/opera/win/605/gd/ Sabhal Mór Ostaig, An Teanga, An t-Eilean Sgitheanach IV44 8RQ Caoimhín Ó Donnaíle, caoimhin@smo.uhi.ac.uk; www.opera.com

[38] DART was a consortium consisting of EBLUL, Sabhal Mòr Ostaig, Fiontar (Ireland), Ofis ar Brezhone.g. (Brittany), Trinity College (Wales) and Opera Software ASA (Norway) with the support of DG Information Society of the European Commission.

Name	Description
Penfriend	Penfriend Ltd is an Edinburgh-based company that provides software tools for disabled people to ease their use of computers. It is sold as a tool for people with dyslexia, cerebral palsy, MS, etc. Products include a screen reader, word prediction, speech feedback and on-screen keyboards for a range of languages. The word predictor is said to reduce the number of keystrokes in typing by c.75%. It works across the vast majority of programs under Windows. In conjunction with Stòrlann and LTS it released a Gaelic lexicon for use with Penfriend containing 10,000 words (with smaller versions containing 500, 2000 and 5000 words each) in February 2007. By comparison, the English lexicon contains 30,000 words and most likely contains a higher number of root lemmas[39] overall due to the limited morphology of English. The Gaelic "package" does not contain a Gaelic voice at the moment although Penfriend does come with several English voices, including Heather (a Scottish voice developed by CereProc, qv). LTS is currently distributing the Gaelic software at reduced rates but pricing is due to increase in the near future. The Gaelic text predictor is currently being used by some Highland schools as an educational tool. The text predictor suggests terms based on the first (and subsequent) keystrokes, the position in the sentence and previously used combinations. It also recognises words without accented vowels. Any of the up to 12 choices may be selected by clicking, using the F keys or the number or numpad keys. The software is also capable of automatically learning new words, in theory an unlimited number thereof. These can also be manually edited in case of mistakes having been entered. CALL Scotland evaluated various supportive writing technology packages in 1999, including Penfriend (version 0.7), gave the product a very good review for effectiveness, combination of features, ease of use and value for money.[40] The company is planning to develop the product's features, including Gaelic and would like to add a Gaelic voice if one became available. It does not presently have the funds or funding in place to drive forward the Gaelic side of the product much. Penfriend Ltd. 30 South Oswald Road, Dùn Èideann EH9 2HG Roger Spooner, rls@penfriend.biz; www.penfriend.biz

[39] "Roots" of a word; for example, the English noun *house* only generates the plural *houses*, whereas Scottish Gaelic *taigh* generates *thaigh, taighe, thaighe, taighean, thaighean*.

[40] See *Features reviews and comparisons of supportive writing technology* in the Attached Files

Name	Description
phpBB	This is an Open Source forum application. The Gaelic version is currently used by Fòram na Gàidhlig (www.foramnagaidhlig.net) and Ionad na Gàidhlig sa Ghearmailt (www.schottisch-gaelisch.de/phpBB2/). Both the localisation of the previous and current versions were carried out by an extremely small number of volunteers. www.phpbb.com
Pools & Pools-T	Pools-T (Producing Open Online Learning System - Tools) is a European funded project to develop software tools that will aid teachers and students in Content Language Integrated Learning (CLIL), specifically including Europe's lesser used languages. It aims to remove some of the burden of educators having to create their own materials. The development teams currently come from Scotland, Greece, Denmark, Belgium and the Netherlands. The tools themselves, however, are aimed at a much larger number of European countries. It also has a large number of cooperating projects such a LANCELOT (the LANguage learning by CERtified Live Online Teachers, www.lancelotschool.com) or the Spanish Audion Project (www.esaudio.net/recordings). POOLS-T projects include the TextBlender which converts a page of text into a web document where all words have been linked to an online dictionary and HotPotatoes and the associated DIY video files which contains a number of tools to produce interactive web-based exercises and games. One of their projects that incorporates Gaelic is Wordlink, a browser-based tool that was developed by SMO (who were one of the chief promoters of the previous POOLS programme). This tool links the words of a web-page to online dictionaries that can then be looked up with a single click. Amongst a variety of other languages this tool is also capable of linking to Gaelic dictionaries. The development of Wordlink began in 2008 and it is currently operational but it could greatly benefit from a lemmatiser and/or lexical database to improve the linking capabilities. It has been released under a CopyLeft[41] license. One additional outcome of the POOLS-T project has been the Guthan nan Eilean (www.smo.uhi.ac.uk/smo/naidheachd/fiosan/guthan-nan-eilean.html) project. Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ Gordon Wells, sm00gw@groupwise.uhi.ac.uk; www.smo.uhi.ac.uk/wordlink

[41] A form of licensing that allows others the freedom to use and adapt resources provided such derivative work is also released under a CopyLeft license.

Name	Description
Pròiseact Comhairle	This one-year project is run by TELI in conjunction with Comhairle nan Eilean Siar and Comhairle na Gàidhlig in Nova Scotia to produce a dictionary of terminology for local government. The project is funded by Bòrd na Gàidhlig and due for delivery in 2010. The scope of the dictionary has not yet been determined and the output will be a digital online file, similar in layout to Faclair na Pàrlamaid without online search functions. TELI, PO Box 1901, Milton Keynes, MK19 6DN Leo McNeir, mcneir@waitrose.com, www.leomcneir.com
Roy Wentworth's spell-checker	Produced by Roy Wentworth. Work-around spell-checker based on an add-on dictionary file for Microsoft Word. Available free of charge for download. Cross-platform. Non-GOC orthography. No exact headword count (manually collated as a custom dictionary in Word), approx. 20,000 word-forms. Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ; www.smo.uhi.ac.uk/gaidhlig/wentworth/litreachadh/

Name	Description
SCOTS	The SCOTS corpus was set up between 2002 and 2007, funded largely by a £300k AHRC grant and went live online in 2004. It covers material from 1945 to the present and is unusual in two respects. The project focussed on requesting donations of digital material by people across the Scots/Scottish English spectrum, including emails, letters, etc. As a result, the project involved little digitisation, although some was carried out. The corpus also contains an unusually high amount of spoken material (and occasional video recordings), some 800,000 (transcribed) words, representing about 20% of the total corpus. The corpus also contains an extremely small amount of Gaelic material. In total, it contains some 4 million words, none of which is tagged. Notable features of the corpus include context search, synchronised audio and map features. The issue of spelling variants was not addressed in the project as the corpus was to be a descriptive rather than a prescriptive resource. The project would like to address this issue (within the corpus) but at present does not have the resources to do so. Existing tools such as VARD (Variant Detector) were found not to be practical for this corpus. The SCOTS project is currently in the maintenance phase and little material is added and no new features are currently being developed. The Corpus of Modern Scottish Writing (CMSW) is the follow-on project from SCOTS and aims to collect a similarly sized corpus for the period 1700-1945 by 2010. This project involves a larger amount of digitisation than SCOTS, as the materials mostly do not exist in digital form. Much of this work is being carried out in conjunction with university departments such as the library's photographic unit. Overall, the team consists of the equivalent of 6 full time staff, with occasional input from other academic staff. At this stage, issues of timing would limit the amount of collaboration that could take place between SCOTS project and a Gaelic corpus project. However, the team would be more than happy to act in a consultative role (similar to their involvement in DASG) and to share expertise with a Gaelic project. There is also interest in potentially making the build of SCOTS more flexible so it could be of use to other corpus projects but this would have to bear in mind the current time and funding constraints as the team is working full-time on the CMSW project. SCOTS, Oilthigh Ghlaschu, 6 Gàrraidhean an Oilthigh, Glaschu G12 8QQ David Beavan, d.beavan@englang.arts.gla.ac.uk; www.scottishcorpus.ac.uk

Name	Description
Stòr-dàta	The online version of the Stòr-dàta Briathrachais is based on the printed version, originally envisaged to be a series of books. An online version was produced in 1994 and the project has been officially dormant since. Some staff at SMO have nonetheless dedicated personal time to expanding the online version. It is currently the only sizeable online source of terms that includes modern terminology and that is available English to Gaelic. It marks parts of speech but no other information is given. Caoimhín Ó Donnaíle, responsible for IT services at SMO, has devoted much of his personal time to developing the online version further. The current version contains additional entries from a variety of sources. The following improvements are currently considered high priority: data-cleansing of the Stòr-dàta database, adding some of the entries missing from the digital version but contained in the print version and dealing with the Leabhar nam Molaidhean (a list of user-submitted suggestions of missing terms). Sabhal Mór Ostaig, An Teanga, An t-Eilean Sgitheanach IV44 8RQ Caoimhín Ó Donnaíle, caoimhin@smo.uhi.ac.uk; www.smo.uhi.ac.uk/gaidhlig/faclair/sbg/lorg.php
TELI	TELI (The European Language Initiative) is a small non-profit linguistic consultancy run by Leo and Cassandra McNeir. Leo McNeir's background is in crime-writing and education, first with the Inner London Education Authority, then with the Institute of Linguistics. Since 1992 TELI has been the language advisory body for the European Association of Local Government Chief Executives (UDiTE, www.udite.org), with which he continues to be loosely associated. From 1993 onwards, TELI was involved in the production of various dictionaries, mainly for local governments. In 1999 TELI produced a dictionary for the Welsh Assembly (with a revised edition published in 2000) and in 2001 launched Faclair na Pàrlamaid. TELI has been actively involved in numerous Gaelic projects and one Irish project since. TELI operates on the basis of inviting (paid) experts with technical expertise deemed necessary for a given project. Some of the materials produced by TELI have been controversial. The Welsh Assembly dictionary, for example, was criticised for a lack of terminology standardisation according to international standards; and for using manually typed Word documents rather than a computerised database for the creation of the dictionaries, leading to inconsistencies between the English-Welsh and Welsh-English versions. There were similar criticisms of the Faclair na Pàrlamaid for errors, inconsistencies, ambiguities and variation. TELI, PO Box 1901, Milton Keynes, MK19 6DN Leo McNeir, mcneir@waitrose.com, www.leomcneir.com

Name	Description
TM	UHI are currently investigating the acquisition of a CAT tool, possibly including a content management system. It has reserved a budget of £20,000 to acquire/set up the CAT tool and a further £5,000 annually for maintenance. The project is in the early stages of planning with a delivery date of December 2010. Although Open Source TM software was considered, UHI placed an invitation to tender with proprietary suppliers in August 2009. The TM project is part of UHI's general intention to create a bilingual technology centre. Institiùd OGE nam Mìle Bliadhna, Oifis-Stiùiridh, Slighe Nis, Inbhir Nis IV3 5SQ Regarding the TM: Ruairidh MacAoidh, ruairidh.mackay@uhi.ac.uk Regarding general plans: Anna Nic an Fhùcadair, anna.walker@uhi.ac.uk
T-Rex	T-Rex is a project to produce a Gaelic thesaurus project and was launched in 2007. It is run by TELI and based loosely on An Dearbhair. Output is planned to be as a digital online file. Launch is planned for the latter part of 2009 and the project is funded by Bòrd na Gàidhlig, LTS and HIE. TELI, PO Box 1901, Milton Keynes, MK19 6DN Leo McNeir, mcneir@waitrose.com, www.leomcneir.com
TTS (J Berry)	Based largely on the 1997 project by Maria Wolters to produce a diphone-based system, Jeff Berry, a PhD student at the University of Arizona, presented another attempt at a diphone system in 2008. In Phase 1 of the project the Wolters diphone set was used and a native speaker was recorded reading a prepared text. The data was then labelled and a limited pronunciation dictionary created for use with Festvox (a speech synthesis system, http://festvox.org/). The prototype can currently only handle an extremely limited set of words. According to the documentation, Phase 2 would require (amongst other things) new recordings, revision of the diphone set, a grapheme to phoneme conversion system to enable the system to deal with items not in the pronunciation dictionary and training an HMM-based[42] system. Unfortunately Mr Berry was unavailable for an in-depth discussion. Jeff Berry, jjberry@email.arizona.edu; http://yllab.dyndns.org/~group2

[42] Hidden Markov Model (HMM), a new type of speech synthesis technology still under development.

Name	Description
TTS (M Wolters)	In 1997 Maria Wolters submitted a PhD thesis on the topic of producing a diphone based TTS system for Gaelic. The system designed consisted of a text to phoneme engine and Festival (www.cstr.ed.ac.uk/projects/festival/), a speech synthesis module developed by Edinburgh University. The system was based on the dialect of Bayble and included a c.2,000 word pronouncing lexicon. Virtually all data from the project (LEUGH), except for the thesis, have been lost in the intervening period. Maria Wolters describes the end-product as "not saying altogether that much" and is firmly of the opinion that sole diphone systems now represent obsolete technology and that a unit selection system should be chosen in preference. Centre for Speech Technology Research, 2 Buccleuch Place, Dùn Èideann EH8 9LW Maria Wolters, mwolters@inf.ed.ac.uk; www.cstr.ed.ac.uk
TTS (Murray & Black)	Iain Murray, together with Mòrag M Black (a PhD student at the time), worked on a prototype for a Gaelic TTS system in 1993. The project used a basic set of text to phoneme conversion rules and the LSI Phonetic Synthesiser (an English engine) to mimic the sounds of Gaelic to avoid having to record and analyse Gaelic speakers, resulting in an underlying English accent. The output was deemed "acceptable". The project was later abandoned due to lack of interest in the wider Gaelic world and a lack of funding. Dundee University, School of Computing, University of Dundee, Dùn Dèagh DD1 4HN Dr Iain Murray, irmurray@computing.dundee.ac.uk; www.computing.dundee.ac.uk

<table>
<tr><th>Name</th><th>Description</th></tr>
<tr><td>Ubuntu</td><td>Ubuntu Launchpad is an online translation project that relies on volunteers from the internet community to provide translations. It provides a simple online translation platform for the Open Source operating system Ubuntu. It also contains a range of other associated Open Source applications such as the Firefox web-browser and the Thunderbird email application.

Translation projects exist for all Celtic languages but are at different stages. The table below gives a comparison of the stages of translation (French and Hungarian have been added for comparative purposes); figures as of May 2009:

<table>
<tr><th>Language</th><th>Total Strings</th><th>Untranslated Strings</th><th>Firefox Completed (1,929 strings in total)</th></tr>
<tr><td>French</td><td>436,586</td><td>42,865</td><td>100.00%</td></tr>
<tr><td>Hungarian</td><td>436,586</td><td>126,415</td><td>100.00%</td></tr>
<tr><td>Irish</td><td>436,586</td><td>287,664</td><td>100.00%</td></tr>
<tr><td>Breton</td><td>436,586</td><td>292,092</td><td>100.00%</td></tr>
<tr><td>Welsh</td><td>436,586</td><td>337,858</td><td>99.00%</td></tr>
<tr><td>Cornish</td><td>436,586</td><td>435,881</td><td>10.00%</td></tr>
<tr><td>Gaelic</td><td>436,586</td><td>436,102</td><td>0.50%</td></tr>
<tr><td>Manx</td><td>436,586</td><td>436,577</td><td>0.30%</td></tr>
</table>
The current project seems to suffer from the “usual” problems: lack of QA, few regular (fully fluent) contributors, no agreed termbase, etc.

https://translations.launchpad.net/ubuntu/jaunty/+lang/gd</td></tr>
<tr><td>University of Abertay</td><td>The School at Abertay is involved in a number of SALT projects but none of them currently involve Gaelic. Dr Kenny McAlpine who has a special interest in digital heritage, digital archival and restoration work could also prove to be a useful contact in relation to the digitisation of older material.

School of Computing, University of Abertay Dundee, Bell Street, Dùn Dèagh DD1 1HG
Dr Kenny McAlpine, k.mcalpine@abertay.ac.uk; www.abertay.ac.uk/Schools/CAT/</td></tr>
</table>

Name	Description
University of Dundee	The School of Computing at Dundee is involved in a number of SALT projects but none of them currently involve Gaelic. Most of these focus on assistive technologies for people with disabilities such as language development systems for children and augmentative and alternative communication. Certain types of technology may have applications in other fields too and some researchers have Gaelic or more generally SALT-related interests. School of Computing, University of Dundee, Dùn Dèagh DD1 4HN Dr Iain Murray, irmurray@computing.dundee.ac.uk; www.computing.dundee.ac.uk
University of Edinburgh	Edinburgh University's Centre for Speech Technology Research (CSTR) is generally considered to be one of the leading centres of SALT in the world. Their work broadly splits into two main branches, Speech Synthesis and Speech Recognition. They have been involved in the creation of voices for various languages and are involved in cutting-edge speech recognition R & D. Their primary interest lies in projects that further the understanding of speech-related topics. They also welcome knowledge transfer projects. CSTR, University of Edinburgh, Informatics Forum, 10 Crichton Street, Dùn Èideann EH8 9AB Simon King, simon.king@ed.ac.uk; www.cstr.ed.ac.uk
University of St Andrews	Most projects at the university currently do not appear to touch upon SALT. However, the university is involved in a corpus project for Egyptian texts. This includes the inclusion of interlinear texts for users, posing the challenge of finding a way of coordinating texts with variant transcriptions and interpretations in a user-friendly way. University of St Andrews, North Haugh, Cill Rìbhinn, Fìobha KY16 9SX Mark-Jan Nederhof, mjn@cs.st-andrews.ac.uk
Web 2.0 Applications	Cànan has been employed by Highland Council to produce web-based teaching and learning resources. The aim is to help passive speakers, semi-speakers and lapsed speakers to achieve fluency and to create a learner community network spanning Argyll & Bute, the Western Isles and Highland Council. The fundamental idea is to use Web 2.0 technology (online based interactive content) to enable users to learn at a time and speed of their choosing. The project is currently in its initial research phase and no concrete plans have been produced to date. Cànan, Sabhal Mòr Ostaig, Slèite, An t-Eilean Sgitheanach IV44 8RQ Flòraidh Forrest, floraidh@canan.co.uk; www.canan.co.uk

Name	Description
Wikipedia	The Gaelic version of Wikipedia, an open online encyclopaedia, currently has 7,176 articles. By comparison Irish has 8,654; Manx 2,815; Welsh 23,871; Breton 27,929; Cornish 1,783; English 2,976,299; and German 936,900 (as of July 2009). Most Gaelic articles are rated "stub-class" (i.e. they are extremely short, some of them consisting only of a single sentence) and there are less than a dozen regular contributors. http://gd.wikipedia.org/wiki/Prìomh-Dhuilleag
Will Lamb	This corpus project was part of Will Lamb's 2002 PhD thesis at the University of Edinburgh, *A corpus-based Study of Scottish Gaelic Speech and Writing*. It is a written corpus, containing approximately 82,000 tokens. It consists of two sub-corpora, one consisting of the transcription of approximately 42,000 words of spoken Gaelic, uniquely containing approximately 11,000 words of informal spoken Gaelic: ▪ Spoken sub-corpus: conversations, radio interviews, sports broadcasts, traditional storytelling ▪ Written sub-corpus: academic prose, fiction, popular writing, radio news scripts It is a tagged corpus, meaning the tokens have been marked with some 100 different tags for linguistic aspects such as clefting, subordinate clause types, code-switching, types of clauses and types of possession. The software was custom-built by Will Lamb's brother under the working title LinguaStat. It was programmed for a pre-Windows 2000 platform under Visual Basic. Since the publication of his PhD, Will Lamb has been unable to continue work on the corpus due to time constraints and problems with running the corpus software. It was written for an early version of Windows and needs recompiling before it can run on Windows 2000 or later. He is interested in developing the project in some way. Alistair Lawson of Napier University has approached Will Lamb regarding the corpus but, to his knowledge, no specific steps have been taken to date. Colaiste a' Chaisteil, Làrach Beinn na Faoghla, Lìonacleit, Beinn na Faoghla HS7 5PJ Will Lamb, will.lamb@lews.uhi.ac.uk

Table III - Index of Proposed Gaelic Projects

Name	Description
Faclair Bun-tùsach	TELI applied for funding from Bòrd na Gàidhlig in 2008 for a four year project to produce a concise English to Gaelic dictionary, containing approx. 50,000 headwords. Some preparatory work has been done but no agreement has been made to date (June 2009). TELI, PO Box 1901, Milton Keynes, MK19 6DN Leo McNeir, mcneir@waitrose.com; www.leomcneir.com
Grammar-Checker	James Galbraith has been in contact with the SFC, SMO and the University of Aberdeen regarding the possibilities of developing a Gaelic grammar-checker. To date there are no concrete plans or agreements. James Galbraith, 53 Davidson Way, Dùn Èideann EH54 8HQ; james.galbraith@scottish.parliament.uk
Talking Dictionary	A proposal by Marc Farr of the North Highland College to produce an online dictionary capable of handling sound to aid learners with pronunciation. Currently not on official College proposal. Mr Farr has indicated interest in collaborating with the AFB project. The North Highland College, Main Centre, Ormlie Road, Inbhir Theòrsa KW14 7EE Marc Farr, marc.farr@thurso.uhi.ac.uk
CereProc	CereProc, founded in 2005, is a Scottish company specialising in text to speech technology. It has strong links to Edinburgh University and is regarded as one of the world's leading developers of TTS technology. More attention has recently been devoted to developing more natural sounding voices, tackling emotion and reducing the cost involved in building a voice. CereProc is also one of the first companies that embraced regional accents in speech synthesis and has developed Scottish and English regional voices. The company has an interest in developing cutting edge technology for lesser resourced languages. As a Scottish company, they have also stated that this "naturally" includes Gaelic. Although the company has worked with a number of large and medium sized languages, to date they have not had the opportunity to do so with a small language and are keen to develop a showcase for others. To create a state of the art Gaelic voice using unit selection (rather than the older diphone system) would require an estimated 25 hours of recording of a native speaker. It is possible to work with less material but since the degree of naturalness and the overall quality is incremental, "more is better". This material, along with analysis of the phonological and linguistic structures, forms the basis for the "front end technology". Working off such a dataset, in contrast to diphone systems for example, also future-proofs the technology.

Name	Description
	Developing the front end involves establishing text to sound rules. A language of course should not be expected to change to suit technology. However, as Gaelic is not fully codified as yet, future codification should bear in mind that increasing orthographic ambiguity (such as the creation of homographs by merging *'nan* and *nan*) also results in computational cost and reduces the quality of output. A new type of technology CereProc is involved in driving forward is Parametric Synthesis which, based on an existing voice, can create additional voices with particular accents on the basis of extremely little material (in the region of 1-2 hours of spoken material). The quality of the outcome is not as crisp as that of a voice built using the previously described method and (currently) does not improve incrementally with more audio material being used. As it requires extremely little material, however, it could have applications in Scotland for producing new material of moribund dialects or dialects only preserved on recordings. CereProc has also created text to speech systems for the Scottish Examination Board to aid students with dyslexia. Monolinguistic systems are naturally possible but best practice is the creating of a bilingual model in bilingual settings such as the same voice for both Catalan and Spanish. This also enables the TTS to handle non-native words, phrases or sentences without interrupting the flow of the voice and changing accent. The immediate application of such a Voice would be within software such as screen-readers and dialogue systems but are by no means limited to interfacing with software. Potential applications could be found in teaching and homes of children in GME with no Gaelic speaking parent. CereProc also aim to have their technology functioning on a number of top range mobile devices before the end of the year (iPhone, LG). CereProc, Appleton Tower, 11 Crichton Street, Dùn Èideann EH8 9LE Dr Matthew Aylett, matthewa@cereproc.com; www.cereproc.com
Grammar Dictionary	A proposal by Iain Fraser Grigor to put together a small team (himself and an editor) to produce an online dictionary of grammar to explain basic and advanced aspects of grammar in relation to Gaelic. The work would be based on existing works (which we already note are inconsistent/incomplete and require further detailed professional research before they are usable). It is estimated that Grigor's proposed Grammar will take 6 months at a cost of approximately £60,000. Mr Grigor holds a BA, MPhil and DipEd from Jordanhill and has studied Gaelic grammar for a number of years but is not a fluent Gaelic speaker. Iain Fraser Grigor, Dunrui, Mòrar, PH40 4PA; iain-fraser-grigor@hotmail.co.uk

Table IV - Index of Projects in Other Countries

Note that the focus in terms of Basque projects has been on the Autonomous Community of Euskadi, as the majority of development is both historically and currently found in that region of the Basque Country. It largely ignores the French Basque Country and Navarre.

Name	Description
Abair	The Irish speech synthesis project, Cabóigín I,[43] is based on WISPR (Welsh and Irish Speech Processing Resources), an earlier research project carried out between the Canolfan Bedwyr, TCD, DCU and the ITÉ.[44] WISPR was funded by a £221,097[45] European Interreg IIIa grant from 2003-05 and produced an annotated speech corpus for Irish and established a network of experts. Subsequent funding has come from Foras na Gaeilge. The first beta Voice (referred to as Cabóigín I) is based on a Donegal speaker, of whom some 10-15 hours of recorded material form the basis of the Voice. Version 1.0 is estimated to be ready for launch some time in 2009/10. The team is currently also working on a Connacht Voice and will commence work on a Kerry Voice soon thereafter as part of the project (now called Cabógaí). Abair has also produced a Firefox plug-in which allows users to synthesise Irish words and phrases while reading a web-page. It is expected that as the software is refined, it will become an increasingly useful tool in teaching, learning and for members of the Irish-speaking disabled community. In terms of planning, the development of a speech synthesis centre of excellence at TCD was described as “mostly accidental” rather than the result of strategic planning. TCD had considered going with a commercial partner initially but decided against it to foster an indigenous and local skills base (which overall are not common at this academic/scientific level), to be able to guarantee long term maintenance and commitment and to foster “ownership” of the project. In the director’s words “such a project is about more than a voice”.

[43] Donegal Irish for “chatterbox”.

[44] The Institiúid Teangeolaíochta Éireann, closed in 2003.

[45] With a total project cost of £294,797.

Name	Description
	The current system is based on the Open Source Festival engine (www.cstr.ed.ac.uk/projects/festival/) which is not considered industry standard anymore, nor does it make an ideal solution for live screen readers for example. There are also issues with the installation of Festival based systems. However, TCD is currently actively working on a more compact solution in conjunction with industry experts, which will produce (free) industry standard voices that will be compatible with screen reading technologies, etc. This new engine will be a hybrid HTS/HMM[46] solution to ensure a balanced output in view of the shortcomings of each individual system on its own. In line with current developments in the industry, the team is also working on building voices capable of emotion. The project's principal investigator herself, Ailbhe Ní Chasaide, has a professional interest in the linguistics of emotion and the uses of speech synthesis in accessibility and education. Although the initial project focussed on Irish and Welsh, it has always been the explicit goal of the Irish side to include Gaelic and Manx at a future date. TCD is also keen to facilitate developing the local Gaelic skills base for speech synthesis. The recommendation would be to record both a unit selection and diphone corpus at the same time to ensure that the engine is capable of dealing with the "gaps" in the unit selection corpus. Recording the data for both in an appropriate format and preparing it accordingly should future-proof the material against any technological advances. In terms of costs, Abair currently is on a £90,000 p/a grant from Foras na Gaeilge and estimates that a joint Gaelic project with TCD would require roughly as much per year, over a period of 2 (ideally 3 years) to produce the first voice. Subsequent voices would be quicker and cheaper to build. This would likely involve a TCD-based collaborator and a Scottish-based collaborator with a sound understanding of Gaelic phonology and recording technology or alternatively a sound technician. Given the linguistic proximity of the two languages, TCD envisages that a lot of the groundwork already carried out could be "tweaked" to suit Gaelic and reduce the need to start developing tools from scratch. An tSaotharlann Foghraíochta agus Urlabhra, Foirgneamh na nEalaíon, Seomra 4091, Colaiste na Tríonóide, BÁC 2 Prof. Ailbhe Ní Chasaide, anichsid@tcd.ie; www.abair.tcd.ie

[46] Simplfied, HTS is a blend of a diphone and a unit selection system. Again simplified, diphone systems are reliable but unnatural. Unit selection systems are natural sounding but slower and with gaps, sometimes causing it to make glaring errors. An HTS hybrid systems "combines" the best of both systems.

Name	Description
Acmhainn	Acmhainn is an online terminology database run and maintained by Traslán (q.v.). It contains most of the terminology published by An Gúm/An Coiste Téarmaíochta and came together following a number of translation workshops funded by Foras na Gaeilge in 2001. It was officially launched in 2002 and funded until 2006 by Foras na Gaeilge. Since then, it has been funded by Traslán. Through Traslán's involvement in the Focal project, new terminology added to Focal is also added to Acmhainn. For historical reasons the older terms are not stored in database format so the number of terms cannot be ascertained fully but is estimated to be in the region of 100,000 entries. *See Traslán for contact details*

Name	Description
An Coiste Téarmaíochta	The primary roles of An Coiste Téarmaíochta are to: ▪ Develop and provide standardized terminology for use by the education sector, the State and the wider Irish-speaking community ▪ Facilitate lexicography for the Irish language using modern working methods and means of maintenance and distribution Historically terminology first began to be developed on a large scale within the education system from the 1930s onwards. This included the use of language expert committees. An Coiste (then *An Buanchoiste Téarmaíochta*) was not set up as an entity until 1968 by the then Minister of Education. An Coiste has a terminology committee of 20 who convene on a monthly basis. Originally comprised of academics, the committee now consists of trained terminologists who agree principles, manage the workload through subcommittees and sanction terminology. The (smaller) subcommittees themselves consist of subject experts who deal with terminology within their subject areas. Originally An Coiste's output was published in specialised dictionaries (e.g. on computing, biology, business, etc.) but the output is now increasingly available as digital-only. To date, most committee members are volunteers but it has been found that paid teams provide better value for money as the expectations (on both sides) are much more clearly defined. An Coiste also provides terminological services to commercial companies. For example, it has worked together with Microsoft on the Irish localisation projects. In the case of Microsoft, Foras na Gaeilge co-ordinated the 2004 Community Glossary[47] project which usually predates a language pack and/or localisation project. Crucially, the terminology developed for projects such as Microsoft is now part of the Focal database. 24-27 Sráid Fhreidric Thuaidh, BÁC 1 Fidelma Ní Ghallchobhair, fnighallchobhair@forasnagaeilge.ie; www.acmhainn.ie

[47] A Community Glossary is a basic list of terminology that is collated and debated by speakers of a given language for future use in software localisation by Microsoft. Oddly, although such a glossary was created for Scottish Gaelic, it is not listed on the MS Community Glossary site (www.microsoft.com/language/wincg/).

Name	Description
An Gúm	The primary role of An Gúm, since 1926, has been the publication of Irish books. Originally part of the Department of Education, it was merged with Foras na Gaeilge in 1999. To date it has published more than 2,500 books in Irish. Today its main focus remains the publication of Irish books, in particular for the education sector, children and dictionaries. As such, it is involved in various projects such as the New English - Irish Dictionary (see New Corpus for Ireland). 24-27 Sráid Fhreidric Thuaidh, BÁC 1 angum@forasnagaeilge.ie
Bwrdd yr Iaith Gymraeg	The Welsh Language Board (Bwrdd yr Iaith Gymraeg) was established in 1993 under the Welsh Language Act, charged with the promotion of the language, the facilitation of the use of Welsh, the preparation and monitoring of language schemes. It superseded the non-statutory advisory Welsh Language Board which had been set up in 1988 to advise on Welsh language matters, in particular the government. Earlier efforts were mostly concerned with the establishment of a language board, legislation, education, etc rather than developing a SALT strategy. The 1995/96 Information Technology Committee was the first initiative by the Bwrdd to look into the wider role of SALT but it featured little in policy until after 2000. The first main report making reference to the strategic role of SALT in the promotion of Welsh were the 2004 report *Machine Translation and Welsh: The Way Forward* and the 2005 strategy report *The Future of Welsh: A Strategic Plan*. The Plan identified several key areas for urgent development such as ▪ Development of language tools ▪ National database of standardised terms ▪ Developing the Welsh translation sector (including a strategic plan) ▪ Developing the SALT sector (including a strategic plan) ▪ Developing the corpus (including a strategic plan) However, this of course does not imply that no Welsh SALT were being developed prior to that, simply the lack of an overall strategy. By 2005, for example, the Bwrdd had supported the translation of Microsoft interface packs and the translation of OpenOffice and the Canolfan Bedwyr had embarked on a Welsh speech synthesis project. Since then various reports, guidelines and strategy documents that lay out the methods and approach the Bwrdd will take and promote in relation to Welsh SALT have been published. Most notable are:

Name	Description
	▪ *Information Technology and the Welsh Language,* 2006 (q.v.) ▪ *Bilingual Software Standards & Guidelines*, 2006 (q.v.) There have also been numerous reports into more specific aspects such as: ▪ *Promoting the Use of Welsh Technology in Gwynedd and Conwy*, 2008 (q.v.) ▪ *Standardizing Welsh Place-names: Principle and Example*, 2009 (q.v.) **Standardisation of Terminology & Welsh National Database of Terms (www.e-gymraeg.org/bwrdd-yr-iaith/termau)** The first main initiative in this area was the 1995 Panel for Official Welsh which looked into matters of clarity and consistency and made recommendations to the Bwrdd following wide consultation in 1996. The chief recommendation, the establishment of a Department of Language Standards under central government was not successful. This resulted in the Bwrdd using its own funding to set up their own projects of terminology development, including a framework for future projects. In 2001 the responsibility for the standardisation of place-names shifted from the Assembly to the Bwrdd, which in collaboration with the Canolfan Bedwyr, Ordnance Survey, the Place-names Research Centre and Powys Council drew up guidelines for standardisation of place-names. The place-names team and the standardisation body were merged into the Corpus Planning Unit in 2003. The National Database of Standardised Terms, developed by the Canolfan, was launched in 2005-2006 and includes both terminology developed by the Canolfan and the Bwrdd. The data can be accessed online but also downloaded as a TM for use in translation software. As part of a project in 1998 with the Canolfan Bedwyr, the Bwrdd set up a terminology panel. This was later expanded and formalised into the Terminology Standardisation and Translation Unit, which aims to become the national coordinator of terminology standardisation in the future. **Software** Although the Bwrdd had given some minor grants in support of the development of the Cysill spelling and grammar-checker (see Cysgliad), its first main involvement in the development of SALT was the collaboration with Microsoft, the translation company Cymen (www.cymen.co.uk), the software company Draig (www.draig.co.uk) and the Canolfan (to handle terminology) to produce a Language Interface Pack for Windows XP and Office 2003. The latest versions currently available are for Windows Vista and Office 2007 and these are available to the public free of charge.

Name	Description
	It has also provided financial support to the Mercator Centre for a localisation project of OpenOffice (see Open Source projects). Beyond that, however, most Open Source projects continue to be run by volunteers. The Bwrdd states that is "open to requests for help" but has no specific strategy regarding Open Source projects. It does state that it will continue to engage with other software developers to investigate the possibility of developing the Welsh software market. Bwrdd yr Iaith Gymraeg, Siambrau'r Farchnad, 5/7 Heol Eglwys Fair, Caerdydd CF10 1AT Lowri Williams, lowri.williams@byig-wlb.org.uk; www.byig-wlb.org.uk
Canolfan Bedwyr	The Canolfan Bedwyr is a Welsh language technology centre, part of the University of Bangor in North Wales. Since its inception it has been at the forefront of Welsh SALT and lexicography and has, amongst other things ▪ Produced a spell-checker ▪ Produced a grammar-checker ▪ Developed Welsh speech synthesis software ▪ Been at the centre of the standardisation of Welsh technical terminology It goes back to 1993 as the vision of a small group of people and was initially associated mostly with the Education Department. In its initial period, it consisted notionally of a loosely associated group of people and projects formally spread across various departments. By 2001 it had by far outgrown the Education Department and, in a period of amalgamation of several colleges, the Canolfan Bedwyr set up as a distinct unit. It was set up with a very broad remit to cater for the needs of Welsh within the university, including translation and a dedicated officer to police the Welsh language policy of the university. Within itself, the centre also has no particular overarching strategy except that for each project, the appropriate skill set is called upon, irrespective of the Welsh speaking skills of the specialist in question. It has today transformed itself into a stand-alone centre of excellence for Welsh speech and language technology and terminology. **Terminology & Centre for Terminology Standardisation (www.termau.org)** The development of terminology in Wales was largely "organic" in nature in the sense that there was no central plan that called for the development and standardisation of technical terminology. However, the 1967 Welsh Language Act did have the indirect effect of creating a need for such a development.

Name	Description
	Delyth Prys, a lexicographer by training, played a crucial role in setting up the Canolfan and its role in terminology. She began her association with the University of Bangor through the then new National Curriculum which had created a need for a standard dictionary for schools. In addition to the dictionary, ACCAC (now merged with the Department for Education Lifelong Learning and Skills (DELLS)) also required a computerised database. The Termiadur Ysgol 5-16 project (a dictionary for schools containing 35,000 terms) had at its core a lexicographer and a terminologist working together and crucially based their methodology on international best practice (the ISO standards on terminology standardisation).[48] [49] As part of their methodology, the Canolfan also worked (and works) with Welsh speakers and subject specialists in conjunction to ensure an optimal skill mix for developing terminology. These steps were considered crucial to ensure consistency and quality, as terminological stability was deemed particularly important for Welsh as a lesser-resourced language. As a result, the Termiadur terminology has been stable since its inception. Most of the Canolfan's terminological work since has been based on the termbase, including the new 5-19 dictionary (containing 48,000 terms) and a string of specialist dictionaries. The Canolfan has hired a terminologist on a 2-year contract, due to start later in 2009 to provide expert assistance to the various ongoing terminology projects. The first edition of Y Termiadur Ysgol was published in book-form only but the second edition was also made available free on CD and has been online since 2007, also as a download, including a version for mobile phones. Looking back, the Canolfan's staff consider that having designed the original database properly, including marking parts of speech, etc, as invaluable. This design has since enabled a flurry of larger and smaller projects, not least of all various word games for BBC Cymru. It now contains approximately 150,000 terms. **Cymraeg Clir (Clear Welsh)** The Canolfan was also the initiator of the Cymraeg Clir scheme to promote a register and style of writing in official Welsh documents. The aim of the scheme is to encourage a more natural (from the Welsh point of view) style of language in official documents to avoid Welsh speakers turning to the English versions due to hard to read translations.

[48] The Canolfan has since become a member of the ISO board, representing both the voice of Welsh and that of small languages.

[49] ISO (International Organization for Standardization, www.iso.org); see 639-1, 704, 860, 1087, 1951, 10241, 12199, 12200, 12615, 12616, 12620, 15188, 15836

Name	Description
	Funding In a nutshell, the arrangement between the University and the Canolfan can be summed up as ***"you can do what you deem necessary as long as you find the funding yourselves and as long as it does not bring the University into disrepute"***. This, amongst other reasons, is why Cysgliad (q.v.) is sold as a commercial product. Beyond that, the Canolfan has been very successful in finding funding, in particular by targeting sources of funding not specifically earmarked for the Welsh language. This arrangement has shortcomings. The University takes any surplus funds and as a result, the Canolfan has not been able to build up a financial cushion during good years and had to downsize in 2008 due a financially disastrous year. Due to its official status as a "service department", it is not part of any of the 6 academic colleges, so the Canolfan also has difficulty in accessing information on sources of funding. Looking to the future a more stable financial arrangement would be welcome, including alignment to an academic college. **Collaboration** The Canolfan is extremely keen to collaborate with other Celtic languages for mutual benefit to develop better and more extensive tools than could be developed by each language community in isolation. Canolfan Bedwyr, Bangor, Gwynedd LL57 2PX Delyth Prys, d.prys@bangor.ac.uk

Name	Description
CEG	The development of Cronfa Electrone.g. o Gymraeg (CEG) or Electronic Corpus of Welsh coincided with that of the Canolfan. Funded by a £21k grant from the Higher Education Funding Council for Wales to the Welsh IT Unit and the School of Psychology, at the University of Bangor, CE.G. was set up between 1993 and 1994 in collaboration with the Department of Welsh. It contains over a million words, mainly of post 1970s material. Due to the lack of existing electronic Welsh texts at the time, the need to use OCR technology to digitise texts and the subsequent need to proofread these vigorously played an important role itself in pushing the need and development of the Welsh spell-checker. Using OCR and a modified spell-checker, one full-time and one part-time researcher averaged about 1,000 words per hour. Future plans for the corpus include: ▪ Enlarging the corpus. ▪ Automatically tagging new additions but manually tagging a subset in greater detail. ▪ Tagging texts for original language, native/non-native author. ▪ Creating a "family" of corpora, including monolingual, bilingual and spoken corpora (including transcriptions). ▪ Adding material not normally included in non-specific corpora such as literature, poetry, etc to address the gaps in written Welsh (lack of daily newspapers, magazines, etc.). Properly tagged, these could be in- or excluded as required when searching the corpus. www.bangor.ac.uk/ar/cb/ceg.php.en
CEMLL	The University of Ulster at Coleraine participates in the CETL (Centre for Excellence in Teaching and Learning) government initiative via its CEMLL (Centre for Excellence in Multi-media Language Learning). Through research carried out between 2005 and 2007 by the centre, in particular on the effectiveness of language learning in language labs, the centre identified a need for (better) CALL resources for the learning of Irish. One particular need that was identified (see Teaching for Transition) was the necessity to "level the playing field" for Irish-speaking university students who enter the university with a broad range of Irish language skills (learners/native speakers, varying command of pronunciation/grammar), including disparate theoretical language skills (such as knowledge of grammar, etc.). A programme was developed that runs through Year 1 that aims to advance students' linguistic skills in a consumable manner to a common level. The research (see Multimedia Language Learning in Higher Education in the UK) identified various problems with the traditional approach to language teaching, in particular in regard to the teaching of Irish through "conversation classes". It was concluded that there was a distinct need for task-based learning which requires a wide range of resources which could be made available in

Name	Description
	a language lab. The tools and programme developed focussed on ▪ Individual **and** group work including peer review ▪ Conducting learning and review through the medium of Irish ▪ Encouraging learning/problem ownership ▪ Maximising exposure and interaction ▪ Producing CALL tools for Irish ▪ Use of current media (such as producing Irish sound clips to video clips taken from the web with the original sound removed) ▪ Setting up an Irish data archive and resource unit with technical support in place The setup developed at Coleraine allows Irish tutors to adapt resources for Irish, develop new resources and to receive training. An example of such a CALL resource is Capaill (www.llas.ac.uk/materialsbank/mb049/index.htm), a series of 16 online exercise units for learners of Irish. Overall, the use of CALL and the language lab has resulted in better use of teachers' time and in providing facilities which students are able to use 24/7. Most of the materials developed are for local use only but the centre would be happy to participate in skills transfer projects. The point was also made that such resources should not be seen as "special dispensations" but as vital tools that should be expected in tertiary education institutions teaching languages. CMLL plans for the future include: ▪ Developing more Irish resources ▪ Focussing more on peer-review with a view to teachers becoming learning facilitators ▪ Developing a common syllabus for Irish at the tertiary level that is compliant with the Common European Framework. (along with UCD Galway and QUB) ▪ Working on the development of mobile language labs in conjunction with commercial partners The centre also runs a translation studies module. This includes a taster course and various units on translation memories and translation resources such as Focal and Irish dictionaries on CD such as Wingléacht[50].

[50] A digital edition of Ó Dónaill's 1992 *Foclóir Gaeilge-Béarla* for Windows (www.nuigalway.ie/cs/staff/software/wingleacht.html).

Name	Description
	Costs A language lab already existed at Coleraine and the project costs were split amongst various university departments (including nursing, the library and the language departments). Staff consists of a director, an Irish lecturer/language technician, a project manager and a technician. Most of the Irish resources had to be developed from scratch as, unlike languages like French or German for which a vast array of commercial CALL products exist, there was virtually nothing for Irish. However, there are an increasing number of Open Source tools in this domain that can be freely adapted to suit individual language needs, some of which are used at the CMLL (e.g. HotPotatoes). As there were existing resources in situ, it is difficult to estimate the exact costs of the project. It is estimated that the personnel costs aside, the cost for a 20 workstation state of the art language lab (including hardware, software and associated licenses) would be in the region of £50k if there are no existing resources. Ollscoil Uladh, Campas Chúil Raithin, Bothar an Chró Mhóir, Cúil Raithin BT52 1SA Caoimhín Ó Dónaill, c.odonaill@ulster.ac.uk; www.cemll.ulster.ac.uk
Collins dictionaries	The University of Ulster, together with Collins Dictionaries, has been involved in a number of dictionary projects since the 1990s. The first of these was the 1995 Collins Irish Gem (640 pages), followed by the Collin Irish Pocket Dictionary in 1997 (640 pages). Work continues on a bi-directional dictionary, the Collins Concise Irish Dictionary (150,000 headwords). The dictionaries are based on the Collins' database of English terms, not on a corpus. The last major dictionary prior to the Collins dictionaries had been the 1986 *Foclóir Póca* (based on de Bhaldraithe's 1959 English Irish Dictionary). The emphasis of the Collins dictionaries was on modern layout and language/terminology and thus constituted a major step forward in Irish lexicography. Ollscoil Uladh, Campas Chúil Raithin, Bothar an Chró Mhóir, Cúil Raithin BT52 1SA Gearóid Ó Domagáin, g.odomagain@ulster.ac.uk

Name	Description
Corpas na Gaeilge	The Royal Irish Academy's *Corpas na Gaeilge* is a historical, monolingual and untagged corpus made available in 2004. It covers the 1600-1882 period, the period straddling the end of the Early Modern Irish period and the beginning of the Modern Irish period. It contains 705 texts from a wide variety of documents of the period. This amounts to approximately 1.2 million words. It also contains an index of personal and place-names and a reverse index. This corpus is not web-searchable and only available on CD-ROM (currently priced €60). The CD itself is searchable. Acadamh Ríoga na hÉireann, 19 Sráid Dhásain, BÁC 2 Úna Uí Bheirn, u.uibheirn@ria.ie; www.ria.ie
Cynllun Sabothol	Canolfan Bedwyr is also part of the Cynllun Sabothol sabbatical scheme in which Welsh professionals, including teachers, can be released from their job for 3 months to go through intensive training to improve their linguistic skills. Two centres exist for this training, one in Bangor and one in South Wales. There are approximately 12 participants per intake and a distance learning version exists. The website also contains links e.g. to the (commercial) Maes-T termbase; an FAQ on Welsh on computers such as accented characters and changing the document language; and video tutorials on these and associated topics (Welsh only). This CPD scheme aside, the Canolfan is not involved in teaching Welsh as a language. www.cynllunsabothol.org

Name	Description
Cysgliad	The Cysgliad package is sold by the Canolfan as a commercial product and contains a digital dictionary, a spell-checker and a grammar-checker. It can be used as a stand-alone programme or integrated within Word/OpenOffice and runs on Windows and MacOSX. It currently retails at around £50 (incl. VAT). For Cysgeir the original dictionary, compiled in the 1990s, had c.48,000 entries and was compiled as a database dictionary. It is bidirectional and updates automatically via the internet if the Canolfan releases an updated version. It can be searched, contains a rhyming dictionary function and a lemmatiser, which means that entering a conjugated and/or mutated word form will automatically direct the user to the root word. Cysill, the spell-checker and grammar-checker is an extension of said database. It is available as a network version and copes with two different registers, periphrastic (less “complex and formal”) and concise (more “complex and formal”). It also does not only suggest corrections but also points out the grammatical rule that was broken, helping the user identify the nature of an error. Cysgeir’s functions originally started life in the psychology department at Bangor University through a researcher’s interest in the neurological implications of Welsh mutations. The development of the (rules-based) grammar-checker began in 1985 but was not carried out as one coherent full-time project. It currently is only able to deal with adjacent words but work is being done to enable it to cope with larger units of speech. This, however, will require recompiling of the grammar-checker. The Canolfan is also working on an error analyser to be able to prompt users who repeatedly make the same mistake with suggestions on how to improve and the integration of the spell-checker into non-Word-processing software such as graphics software and translation software. This is being done using the Open Source HunSpell spellchecking software in conjunction with the company Semantise in a Knowledge Transfer Project (KTP). www.e-gymraeg.org/cysgliad/

Name	Description
eDIL	eDIL is a digital version of the Royal Irish Academy's dictionary of Old and Middle Irish, with some material from later periods. Originally published in 22 fasciculi between 1913-1976, a compact version was published in 1983 and 1990. The online version was launched in 2007. The online version allows (simple and advanced) full-text searches, has improved legibility, widened access beyond the original narrow scholarly circle with access to a printed version. eDIL is now a purely web-based resource and the current form has enabled the team, in conjunction with the scholarly community, to start work on PacDIL. This is an extended version in which ▪ Errors and inconsistencies in the original publications have been amended ▪ Readings have been updated ▪ Additional/new material is added far beyond the original scope (such as material from scientific magazines e.g. Revue Celtique). The data was coded in XML and in accordance with TEI (www.tei-c.org) standards to ensure compliance with international standards and to futureproof the data. This project is collaborating with the University of Cork who run CELT, an online database of historical Irish texts to link eDIL and CELT. It contains approximately 10 million words of the Middle and Early Modern Irish period and includes translated texts and texts in languages other than Irish. This will support largely scholarly research for people without access to a library with a large Irish section. Ollscoil Uladh, Campas Chúil Raithin, Bothar an Chró Mhóir, Cúil Raithin BT52 1SA Prof. Gregory Toner, gj.toner@ulster.ac.uk; www.dil.ie
Elhuyar	The various branches of the Elhuyar Group, a not-for-profit branch and a commercial branch, working to promote and popularise science and technology through the medium of Basque. Elhuyar is the NFP (and oldest) branch of the group. When founded in 1972, it was originally a cultural association which in 2002 became a foundation. Its main funding streams are members' contributions, public funding and profits from commercial products. Eleka is the commercial branch of Elhuyar and focuses on the development and marketing of language technology and services. It grew out of an initial project with Ixa (q.v.)

Name	Description
	Elhuyar Aholkuritza, the consultancy branch, offers consultancy on a wide range of linguistic matters to organisations and companies in relation to Basque but also other minoritised languages. Most notable perhaps is its expertise in drawing up language schemes for bodies, companies, communities or specific demographic groups to improve the usage of Basque within a given domain and to influence language attitudes. The foundation is a keen collaborator and is a member of various organisations in the field of knowledge transfer, research and development, etc such as ▪ The Knowledge Cluster (www.clusterconocimiento.com) ▪ The AlphaGalileo Foundation (www.alphagalileo.org) ▪ Eusko Ikaskuntza (the Basque Studies Society, www.eusko-ikaskuntza.org) ▪ InnoBasque (www.innobasque.com). It has won various awards for its work in the field, in particular its popular science site Zientzia (www.zientzia.net). Some of the projects it has collaborated on or developed include ▪ Subject-specific corpora ▪ OpenTrad, machine translation software (www.opentrad.org) ▪ Elebila, a Basque/Castilian/Catalan search engine (www.elebila.eu) ▪ Xuxen, the Basque spell-checker (www.xuxen.com). Its main Research & Development focus at the moment is on machine translation, translation memories, term and information extraction and corpus research and tools. Its publishing section, which over the years has published hundreds of scientifically related books and multimedia publications, states that its primary target audience is the general public, young people and specialist publications (educators, researchers, subject experts, etc). It is particularly well known for its series of general and specialist dictionaries, bi- and mono-lingual. Elhuyar, Zelai Haundi Kalea 3, Osinalde Industrialdea, 20170 Usurbil, Gipuzkoa, Spain elhuyar@elhuyar.com; www.elhuyar.org
Euskaltzaindia	The Euskaltzaindia, sometimes referred to as the Academy of the Basque Language, was set up in 1919 as an academic body to research, cultivate and promote the Basque language. It gained recognition as a Royal Academy in Spain in 1977 and as a Cultural Association in France in 1995. Of the institutions studied in this report, it is by far the oldest.

<table>
<tr><th>Name</th><th>Description</th></tr>
<tr><td></td><td>
The Euskaltzaindia's work was interrupted significantly by the Spanish Civil War and WW2 and its primary functions (as seen today) as a regulatory body for the language did not commence in earnest until the 1960s. Developing a standard, today referred to as Euskara Batua (Unified Basque) or simply Batua, began with the agreement on how the spelling and grammar of the language would be standardised at the 1968 Congress of Arantzazu. An important milestone was the 1979 agreement between the nascent Regional Government of Euskadi and the Euskaltzaindia, establishing it as the ultimate authority on questions of standardisation.

Although there had been previous attempts at developing one, throughout history, Basque never enjoyed an accepted, common written standard. Instead various of the seven Basque dialects, referred to as Literary Dialects, had enjoyed the status of the quasi written standard depending on where the majority of the literary output was taking place. Shifting between North and South, this had led to various French or Spanish influenced writing systems. For example, the surname Oiarzabal was variously spelled Oyarzabal, Oiarzabal, Oyarçabal, Oyarccabal and Oyharcabal.

Having agreed a spelling, the body then turned to the morphology of the language. Some differences between the seven dialects had grown so large that speakers from non-adjacent dialects would sometimes resort to French or Spanish for communication. For example, going from West to East:
<table>
<tr><td></td><td>'they are going'</td><td>'they sent them to me'</td></tr>
<tr><td>Bizkaian</td><td>daoaz</td><td>bidali eustezan</td></tr>
<tr><td>Gipuzkoan</td><td>doaz</td><td>bidali zizkidaten</td></tr>
<tr><td>Lapurdian</td><td>doatzi</td><td>bidali zauzkidaten</td></tr>
<tr><td>Zuberoan</td><td>doatza</td><td>bidali zeizgüen</td></tr>
<tr><td>Batua</td><td>doaz</td><td>bidali zizkidaten</td></tr>
</table>
In line with attempts in the 1930s to use Gipuzkoan, the central dialect,[51] as a basis for a unified standard. Gipuzkoan also ended up as the source for Batua forms if no consensus forms could be agreed. In spite of the significant challenges,[52] the Aditz Batzordea (Verb Commission) published a unified set of forms in 1973-77 and by 1979 all other outstanding issues of the
</td></tr>
</table>

[51] Also the province with the highest absolute number of speakers in the 1970s.

[52] Not a trivial matter as the auxiliaries for *to have* and *to be* alone have more than 12,000 different forms.

Name	Description
	standardised spelling and grammar had been agreed upon. These were quickly followed by standardised lists of place-names and settlement names and rules on the adoption and spelling of loanwords.[53] As a compromise solution, Batua was hugely controversial for many years and was not immediately widely accepted. Southerners objected to the use of Northern **h** in the standard spelling, Northerners objected to the use of a Southern dialect rather than classical Lapurdian as the main source - in short, there was something for everyone to dislike. But through its persistent use in education, publishing, the media and the administration, it is now the de-facto standard for Basque speakers world-wide. Although most Basques today either naturally acquire Batua in the education system[54] as children or in adult education, all rules of the standard and authoritative word-lists (including surnames and place-names) can be accessed online or downloaded in their entirety free of charge. To this purpose, the Euskaltzaindia also maintains the online *Hiztegi Batua* (Unified Dictionary, www.euskaltzaindia.net/hiztegibatua). This is not technically a dictionary but the sum of terms that the body has specifically standardised over its history. It currently contains the "historical" 20,000 terms and the Euskaltzaindia is currently working on adding another 20,000 terms. As such, it lists the standard form of an item and also the regional variants that exist for it. Batua has been surprisingly stable, with only minor changes having been made. The Euskaltzaindia has, however, on a number of occasions issued rules on points that had not previously been clarified such as the transliteration of names and terms written in non-Latin writing systems, compounding and hyphenation. It continues its standardisation work, in particular in the fields of personal and place-names alongside its other functions. In addition it continues to be active in linguistic research, publishing and a number of other areas that do not touch upon SALT for the purposes of this report. **Terminology and Lexicography** One of the issues the Euskaltzaindia did not address sufficiently quickly after the development of Batua was the question of new/ standardised terminology and lexicography. Although it published a series of word lists and dictionaries from 1973 onwards on topics ranging from mathematics, commerce and architecture, it did not dedicate enough time and effort in the area. The main exception, or perhaps cause, was the *Orotariko Euskal Hiztegia* (see below), a historical dictionary project. As a result, a multitude of dictionaries was published by an ever increasing number of individuals (including members of the

[53] For example, whether psychology should be spelled *psykhologia* (broadly following the French model), *psikologia* (following the Spanish model) or *sikologia* (following the common pronunciation by Basques).

[54] Every child in the Autonomous Community is at the very least traditionally taught Basque as a subject; the situation in Navarre and the French Basque Country is more complicated.

[55] Basque/Spanish/French/English, plus Latin nomenclature of species.

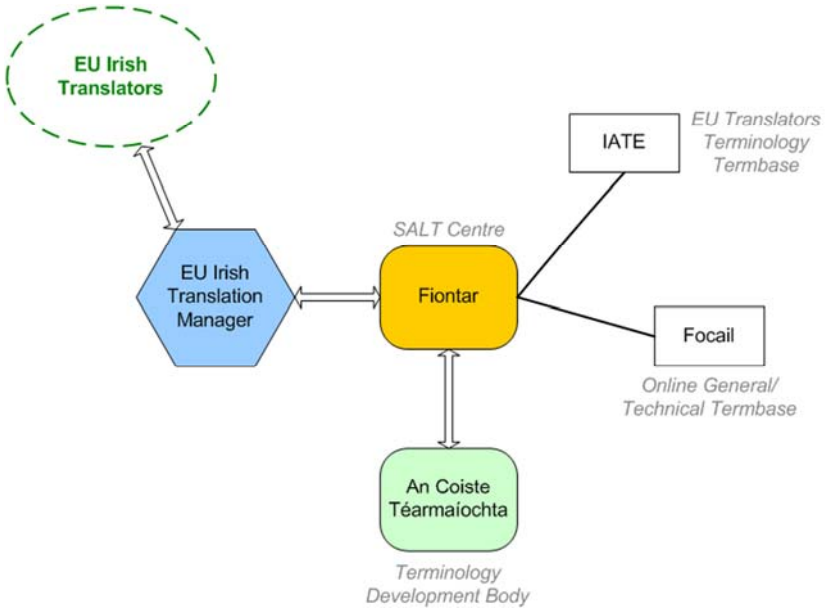
[56] Including a lack of appropriate corpus technology which did not come into existance until the end of the 1960s.

Name	Description
	Euskaltzaindia), groups and organisations. While the principles of Batua were on the whole followed (except for some early disagreements on the treatment of loanwords), the need for new terminology led to an increasingly confusing situation. Depending on the author, new terms were either transliterated from Spanish or French, Basquified to varying degrees by adding Basque endings or created by forming neologisms based on existing Basque words. While Basque, with a myriad of derivative suffixes, is well placed to form new words, many new coinages were too fanciful for the average speaker and never gained wide currency. Efforts by some also continued to "rid" the Basque language of well-established Latin/Spanish/French loans. Two prolific groups stand out, UZEI (q.v.), an organisation dedicated to developing Basque terminology and lexicography. It produced dozens of quadrilingual[55] dictionaries on topics like biology, medicine, politics and printing. The other is Elhuyar (q.v.), a cultural organisation with similar aims. The Euskaltzaindia continues to be consulted on and participates in developments concerning the standard terminology, though today the overall planning and steer is mostly provided by the Language Policy Department (see HPS). One project of particular interest is the planned language industry cluster in collaboration with the Provincial Government of Gipuzkoa to kick-start further developments in the sector. **Orotariko Euskal Hiztegia (OEH, Historical Dictionary of Basque)** A historical dictionary such as the OEH had been part of the Euskaltzaindia's plans as far back as 1918 when the setup of such a body gathered momentum. However, for a wide variety of reasons,[56] work on the OEH did not begin until 1984. The project was supported by the Government of the Autonomous Community, the Government of Navarre and the Provincial Governments. The OEH covers the entire attested period of the Basque language, starting with the Aquitanian material from Roman Gaul until about 1970. Its corpus contains approximately 6 million words. The first volume was published in 1987 and the last, Volume 16, in 2005. The associated corpus is untagged and not publicly accessible. From 1999-2000 the Euskaltzaindia (with the Autonomous Government covering 85% of costs) computerised its processes and the dictionary project to bring it in line with modern developments in lexicography and IT. **Euskararen Corpusa (www.euskaracorpusa.net/XXmendea/index.html)** In collaboration with UZEI, the Euskaltzaindia has also embarked on a corpus of the modern language, the Euskararen Corpusa. It covers the period from 1900 to date and currently contains approximately 4.6 million words. It contains both dialectal and Batua material from over 6,000 publications, virtually the entire stock of 20th-century publications. It also contains transcribed spoken Basque.

Name	Description
	The data is SGML formatted, lemmatized and is capable of recognising variants that are modern dialectal spelling variants or spellings that pre-date Batua (e.g. the Batua spelling *egia* returns instances of *egia, eguja, egiya, eguia*). **EODA (www.euskaltzaindia.net/eoda/)** EODA is an online database of Basque personal and surnames. It is based on research and standardisation work carried out by the Euskaltzaindia. First published as a dictionary in 1972, it is now available online and continuously updated and expanded and currently contains over 10,000 surnames and their modern standard spellings. Euskaltzaindia, Plaza Barria 15, 48005 Bilbo, Bizkaia, Spain info@euskaltzaindia.net
EHU	EHU, the University of the Basque Country (www.ehu.es), is the main public university on the Autonomous Community of Euskadi. It was formed in 1980 by the amalgamation of different campuses in the 3 provinces of the Autonomous Community. It is the main academic research institution in the region. Among relevant projects, EHU runs online degree modules in Open Source Software to increase awareness, understanding and development of OS software. **Euskara Institutua (Institute of the Basque Language, www.ei.ehu.es)** The Institute of the Basque Language was founded in 1996 and is part of the University of the Basque Country (EHU) and more specifically, the Basque Philology Department. Its remit is to study the Basque language from a linguistic point of view. In spite of its name, the philology department does not focus on philological studies alone. Although EHU is gradually increasing the number of degree courses available through the medium of Basque, only about half of the courses can be studied completely through the medium of Basque. To improve the use of technical Basque, it therefore also teaches specific Technical Basque Language modules for students of other disciplines. Indeed, the majority of the courses currently taught fall into this category, for example, Basque for students of Engineering, the Police Force, Chemistry, Architecture, Geology and Environmental Studies. At the Leioa Campus, this is done through the Philology Department. **EPG Corpus** The EPG Corpus is a closed prose corpus containing some 25 million words. Half the material comes from Basque newspaper articles, the other half from prose publications from the period 2000-2007. It ran from 2001-2007 and is being maintained but not added to.

Name	Description
	The sources were carefully selected for their high quality and it was found that for the purposes of prose register research at EHU, adding more material would not significantly increase the amount of information that could be gleaned. Instead, the Institute decided to produce additional topical corpora in the future according to need. Based on this corpus, the Institute has ▪ Carried out research into lexical change ▪ Produced and is producing a number of online topical dictionaries on prose, contemporary language, etc ▪ Developed workshops to improve the teaching through the medium of Basque at tertiary level The EPG Corpus project received financial support from Donostia City Council and the Provincial Government of Gipuzkoa. **ZIO Corpus** The ZIO Corpus is a corpus of translated scientific publications in Basque produced by EHU. As new editions are published, they are also added to the corpus with the aim of furthering research into a developing register of scientific Basque and to support translators working in the field. This project receives financial support from the Provincial Government of Bizkaia. **Aholab** EHU also has a Signal Processing Lab that is currently working on various Basque speech projects. It has to date developed AhoTTS (http://aholab.ehu.es/tts/), an online TTS system for Basque and integrated it into Firefox and Internet Explorer. It is also involved in speech recognition research and development for Basque.

Name	Description
Fiontar	Fiontar is an interdisciplinary centre that straddles Irish, IT and Business/Management. It was set up in 1993 and offers both under- and postgraduate degrees taught through the medium of Irish. It is also involved in various Irish-related IT projects such as the place-names project (see Logainm), the national terminology database (see Focal) and IATE. The initial project, funded by EU Interreg money, was set up with the aim of providing a specialised IT/Business degree for Irish-speaking humanities' graduates. The centre's student numbers are currently down from earlier numbers due to increased competition with the NUI Galway which was identified as the primary future provider of a bilingual campus. Courses on offer are postgraduate degrees in Management and IT (Gnó agus Teicneolaíocht an Eolais) and Bilingual Practice (Cleachtas Dátheangach) and undergraduate degrees in Business & Irish (Gnó agus Gaeilge) and Irish & Journalism (Gaeilge agus Iriseoireacht). Fiontar currently employs 3 full-time teaching lecturers, 3 full-time lecturers focussing on the centre's project work, a number of assistant editors and technicians and a network of external consultants. Ollscoil Chathair Bhaile Átha Cliath, Glas Naíon, BÁC 9 Caoilfhionn Nic Phàidín, caoilfhionn.nicphaidin@dcu.ie; www.dcu.ie/fiontar/index_en.shtml

Name	Description
Focal & IATE	Fiontar has developed the technological side, including management systems, of Focal (www.focal.ie) in partnership with An Coiste Téamaíochta. However, Fiontar is also actively involved in wider terminological work and the European IATE terminology portal.[57] Irish translators to the EU, via a translation manager, send prioritised lists of terms to Fiontar where a team of 4 assistant editors extract existing terminology or partially existing terminology from Focal and on this basis create a list of suggestions which are then validated by An Coiste Téarmaíochta.  **Figure 5 How Fiontar coordinates terminology needs, development and termbases**

[57] Interactive Terminology for Europea, a technical termbase searchable between all 24 official EU languages (http://iate.europa.eu/).

Name	Description
	Fiontar currently expects to be handling some 16,000 terms a year, 10% of which are estimated to be completely new. This follows work on cleaning the original Irish database on IATE. This resulted in a slight drop to currently 14,000 terms but is expected to be on target for the second year target of 30,000 terms. Future terminology plans include the extraction of terms from Irish statutory instruments. This project has been so successful that smaller EU accession languages are now showing an interest in Fiontar's work. Fiontar is also involved in setting up a new group to carry out official translations and formalising the roles of the new body, the Rannóg and An Coiste. Due to a phase of decentralisation, some translation functions were moved away from the Rannóg to various departments, leading to problems with consistency and quality. This has increased the need for re-centralising translation. On the other side, the Rannóg has itself been involved in the creation of terminology, a task mostly carried out by An Coiste, which is also charged with the revision of the Caighdeán (see Rannóg an Aistriúcháin). This has led to a somewhat confused setup that requires clarification and streamlining. The primary function of Focal is to provide a state of the art termbase that is publicly accessible. However, it has also been designed to be used as a management tool for the creation of terminology and ratification. Focal itself started in July 2004 in collaboration with the University of Lampeter under Interreg III and Foras na Gaeilge funding with a total €740,000. The period 2008-2011 is funded by Foras na Gaeilge alone. Prior to 2004 Irish lexicography was not exploiting the available technology to the utmost. Few terminology resources were available online, save for some of the specialist dictionaries published by An Gúm as a online searchable dictionaries on the Foras na Gaeilge-funded Acmhainn site in 2002 (11 dictionaries and 11 terminology lists).[58] Beyond that, there were more than 50 printed dictionaries of various ages and static digital versions of various terminology resources, entailing the usual problems of contradictions, errors, etc. The project involved the digitisation and merger of more than 50 dictionaries and terminology lists in cooperation with An Coiste. It includes terminology not previously published, such as material from the Government's translation service (Rannóg an Aistriúcháin) and the Defence Forces. Not all the material has been edited and approved by An Coiste to date but such items are clearly marked. Some of these terminology lists can be downloaded from the Foras na Gaeilge website as well and An Coiste continues the task of vetting the terminology.

[58] The site is still being maintained and updated by Traslán (q.v.).

Name	Description
	Focal currently contains 315,000 terms and was awarded various prizes, such as the European Commission's European Language Label in 2007. Usage continues to increase. In its first year (2007) Focal handled on average 95,000 searches per month, in 2009 to date the average number of searches per month is 520,000. In the process of developing Focal, the team has developed on-site terminology training for untrained staff. There are plans to expand this training facility and this provides an excellent opportunity for skills transfer. One thing Focal currently does not contain is common vocabulary and enough samples of usage. On the whole it does not cope well with idioms and expressions. Plans for 2009/2011 include: ▪ Samples of usage ▪ Linking Focal to Nua-chorpas na hÉireann ▪ Adding terminology for state bodies, people, countries, etc. ▪ Using the database for the creation of TMs ▪ Publication on CD-ROM *See Fiontar for contact details.*
Foras na Gaeilge	Foras na Gaeilge is the body responsible for the development and promotion of Irish both in the Republic of Ireland and Northern Ireland, set up in 1999 as a cross-border body as a result of the Good Friday Agreement. It also provides funding for a large number of Irish-language projects such as Focal and Logainm (q.v.). In the Republic, it assumed the role of the earlier Bòrd na Gaeilge. The roles of two other bodies - An Coiste Téarmaíochta (The Terminology Committee) and An Gúm (the government publisher of Irish publications) - were also transferred to Foras na Gaeilge in 1999. Although the relationship to the Rannóg and the Royal Irish Academy in terms of lexicographical and corpus development duties remains to be ascertained, the merger of the three bodies has led to much better coordination and planning than previously. For example, until Foras na Gaeilge took on these functions, lexicography was largely an unplanned affair and restricted to isolated projects by individuals or organisations. This led, amongst other things, to the fact that since de Bhaldraithe's 1959 English - Irish Dictionary no other major English - Irish dictionaries had been produced by 1999. Beyond that, Foras na Gaeilge believes that the most influential development in the last decades for the Irish language has been digitisation due to its numerous knock-on benefits; for example:

Name	Description
	▪ Digital termbases and dictionaries have enabled the relatively quick development of follow-on tools and services such as spell-checkers and online termbases. ▪ The digitisation of existing dictionaries and resources has increased the usefulness of existing tools such as eDIL (q.v.), a digital version of the Royal Irish Academy's dictionary of Old and Middle Irish, by enabling text searches, improving legibility, widening access and improving the resource. ▪ An overall reduction in production costs by making lexicographical resources available online, which also reduces the need for re-printing amended versions and enabling immediate correction of errors and addition of material and providing resources which go far beyond the capacities of printed media. The website of Foras na Gaeilge also functions as a hub for people looking for a wide variety of information on Irish. Some examples relevant to SALT issues are: ▪ The Database of Public Sector Terminology where commonly used terminology and phraseology is publicly accessible ▪ The National Terminology Database ▪ An accreditation scheme for translators since 2005 and support services for translators such as regular workshops (Ó Bhéarla go Gaeilge) ▪ A helpline and terminology enquiry form It has been directly and indirectly involved in a number of SALT projects, such as the Irish Language Interface Packs for Windows XP and MS Office, in collaboration with the University of Ulster, Maynooth and Limerick. Foras na Gaeilge currently identifies several areas for development in the general area of SALT: ▪ Digital language teaching aids and digital/interactive pronunciation tools ▪ Using technology to reduce the amount of duplication currently in existence e.g. by providing online templates for official documents ▪ Accreditation of language experts; this includes a planned internal accreditation scheme for editors and revisers in 2010 with accompanying resources[59] Foras na Gaeilge is currently preparing a 20-year plan for the development of the Irish language, due to be published in September. To date, SALT has not been part of dedicated language planning north or south of the border and was mostly a case of "seeing gaps and moving to fill them". However, the new plan may address the issue of SALT planning in general terms. The aim of Foras na Gaeilge is to address needs in terms of SALT without being subject to too detailed plans as technology is unpredictable. Therefore, their primary measures in assessing any such project will likely continue to be whether a need is

[59] Such as the textbooks *Cuir Gaeilge Air!* and *In Ord is in Eagar* (due 2010) produced by Antain Mac Lochlainn, co-editor of www.acmhainn.ie.

Name	Description
	addresses, whether it is value for money and whether it will be used. Foras na Gaeilge, 7 Cearnóg Mhuirfean, BÁC 2 Deirde Davitt, ddavitt@forasnagaeilge.ie; www.forasnagaeilge.ie
Freagra	Freagra is an information service which provides terminological and grammatical advice and very short translations via phone, text and email. The service is free for users and is funded by Foras na Gaeilge and run by Traslán (q.v.). Freagra currently handles some 30-40 inquiries a day, most of which are from various branches of local and national government and administration and to a lesser extent teachers and members of the public. An advertising campaign is planned for later this year. www.freagra.net
HPS	The HPS (Hizkuntz Politikarako Sailburuordetza, Department for Language Policy) sits within the Culture Department of the Government of the Autonomous Basque Community.[60] Although other departments individually have responsibilities regarding language use and promotion, the HPS is charged with the overall task of supporting, promoting and managing the language both within the government and Basque society. Although numerous Basque projects and some planning existed prior to the 1999 General Plan for the Revitalisation of Basque (EBPN), the EBPN was the first to describe a wider strategy to develop this aspect of language development. It specifically identified the following areas for development: ▪ The creation of basic tools fundamental to further developments such as a lexical database, mono- and bi-lingual corpora and the means for morphological, syntactic and semantic analysis of the language ▪ Further development of terminology and better planning thereof ▪ Encouraging the development of proofing tools by the industry, tools supporting translators, terminology resources and speech technology ▪ Developing a Basque IT cluster to localise and develop software ▪ Raising awareness, promoting the topic, providing support, improve networking

[60] Comparable to Arts, Culture and Sports within the Scottish Government.

Name	Description
	Priority was to be given to products that would be of use to a wide audience. Since then, Basque ICT and the need for its development has featured in a number of reports and plans.[61] The recent Basque Corpus Planning (ECP) report documents the then state of affairs. It re-affirms the need for codification (spelling, grammar, lexicon) and the need to elaboration (new terminology, register development) overall and its relation to status planning. It confirmed that virtually all basic issues such as linguistic description (including dialect descriptions across the entire Basque-speaking area), codification (standard spelling and grammar) had been addressed. The issue of appropriate domains for the use of the standard and dialects has also been addressed. Core needs in terms of personal and place-names had also been addressed. It also established that the development of SALT as a whole was progressing well, including (current versions given): ▪ Proofing Tools: Xuxen IV, HunSpell ▪ Office Software: MS Office 2007 (since Word6, in collaboration with Elhuyar), OpenOffice 3.1 ▪ Operating Systems: Windows XP (Interface Pack; since Win95, in collaboration with Elhuyar), EusLinux 2009, Mandrake Linux 2008.1 (project coordinated by the HPS) ▪ Web browsers: IE8 (since IE4), Firefox 3.5 ▪ Management software: SAP 4.6c, Sugar CRM 4.2.0, OpenbravoPOS 2.20 ▪ Plug-ins: QuarkPress parser; Elhuyar dictionaries; UZEI synonym dictionary; OCR 1.1 ▪ Other: WordFast (translation software), PandaAntivirus, OmniPage, School software Collaborative projects and pilots are ongoing in speech recognition (with Telefónica), speech recognition and synthesis (with Scansoft Belgium) and other fields. The departmental website also doubles up as an information portal with exhaustive lists of resources (including Basque software downloads, www.euskadi.net/euskara_soft) and links to relevant external sites. It hosts a number of resources including Euskalterm (see below) and the Toponomastics Database (containing approximately 500,000 place and town-names, 117, 000 of which have been standardised to date).

[61] 2001-2004 Science, Technology and Innovation Plan, 2002-2005 Euskadi Information Sociey Plan, 2003-2005 IT and Telecommunication Plan, etc.

[62] See www.euskara.euskadi.net/r59-4572/es/contenidos/informacion/aurkezpena/es_8550/presentacion.html

Name	Description
	Outlook Apart from expanding existing projects (corpora, terminology, onomastics, etc) and project maintenance, the HPS identifies the ICT sector (including the expansion of Basque on the Internet) as one of the key areas for development. In spite of the small size of the Basque language market, these tools are seen as vital to developing the language. Particular projects identified include: ▪ Expanding the range of Basque software available, both proprietary and Open Source ▪ Continue terminology development, including expansion of the online terminology database and production of some printed dictionaries ▪ Speech technology (the Aditu project[62]) ▪ Machine translation ▪ Bilingual/spoken corpora The general approach will be to ▪ Gradually increase the number of available tools ▪ Concentrate on tools that already exist in Spanish/French/English ▪ Prioritise development according to the needs of users HPS, Donostia Kalea 1, 01010 Gasteiz, Araba, Spain; www.kultura.ejgv.euskadi.net/r46-17893/es/
Ixa[63]	Ixa was set up in 1987 as a research and development group to develop computational tools Basque language users. It is part of EHU (q.v.) and consists of 31 IT specialists, 14 linguists and various other associates. It has to date developed a number of vital basic tools. It has also collaborated with a number of bodies and companies to develop additional tools. These include ▪ EDBL, the Digital Lexical Database of Basque (http://ixa2.si.ehu.es/edbl/), containing some 80,000 words. This database formed the basis of Xuxen,[64] the Basque spell-checker (see below). It can be downloaded free of charge. ▪ Xuxen, a Basque spell-checker. This collaborative development between Ixa and UZEI was financed by the Language Policy Department of the Basque Autonomous Government. It runs on Windows, MacOSX, Linux and can be integrated into various software packages such as MS Office, OpenOffice, QuarkExpress and InDesgin. Currently on Version 4, it was originally released in 1998. It is available free of charge via download. A free online version also exists (www.xuxen.com), financed by the Provincial

[63] Pronounced /i□a/.

[64] Pronounced /□u□en/.

Name	Description
	Government of Gipuzkoa. ▪ Morfeus, a Basque morphological analyser ▪ EUSLEM, a Basque lemmatiser and tagger ▪ Integration of Elhuyar's Basque dictionary into Word, including lemmatising functions that guide the user to the required entry (e.g. when entering a conjugated verb) ▪ MultiMeteo, a multi-lingual weather forecast generator, now also used by the Spanish National Meteorological Service (www.aemet.es/es/nuevaweb). ▪ BertsolariXa; a rhyming dictionary It is currently developing additional tools, in particular to support translators. **The 5 Stages of SALT Development** Based on their experience, Ixa has developed a roadmap to developing language technology, in particular for smaller languages. This roadmap distinguishes 5 main stages:[65] 1. Foundation Stage ▪ Phonetic and morphological description of the language ▪ Basic corpora (spoken and written), not necessarily tagged ▪ Lexical database 2. Second Stage ▪ Statistical corpus analysis, Parts of Speech tagging ▪ Analysers, taggers, generators, lemmatisers ▪ Basic speech processing ▪ Further development of the lexical database

[65] Abridged version, the full version may be accessed on IXA's website

Name	Description
	3. Third Stage ▪ Basic proofing tools (spell-checkers) ▪ Structured (bilingual) dictionaries and integration of text tools ▪ Further development of the lexical database ▪ Study of surface syntax 4. Fourth Stage ▪ Syntactically tagged corpus ▪ Advanced proofing tools (grammar-checkers, style-checkers) ▪ Analysing and developing the semantic level ▪ Integration of tools 5. Fifth Stage ▪ Semantically tagged corpus ▪ Translation aids ▪ Advanced speech processing ▪ Information systems development These stages are not to be seen as completely linear. Certain developments of different stages may indeed run in parallel but the more advanced the technology being developed, the more it will rely on the foundation work of the earlier stages. Adherence to international standards at all stages is deemed critical. The group also encourages the participation of students through a final-year student project. Ixa, 649 Posta Kutxa, 20080 Donostia, Gipuzkoa, Spain acpalloi@si.ehu.es, http://ixa.si.ehu.es/Ixa
Less Mess[66]	Less-Mess is a small software application that provides an on-screen keyboard for a variety of languages, including Irish. The keyboards are limited to the respective special characters appearing in the language. The software, a VisualBasic application, runs under Windows and is said to function in most Windows applications including Office and Internet applications.

[66] Neither Less Mess nor To Bach are necessarily the right approach to the accents issue in Scottish Gaelic but exemplify how such matters can be approached in a simple manner.

Name	Description
	Less-Mess.com LLC, 40 Kietzke Lane, Reno, Nevada 89502, USA greene@less-mess.com^; www.less-mess.com
LinkLine	LinkLine was set up as a helpline for inquiries about Welsh and short Welsh translations in 1998. This service can be accessed by email, phone, fax and sms and was originally operated by postgraduate students at the University of Cardiff. This service was re-launched in 2005-2006 and is now run by the Bwrdd itself. http://www.byig-wlb.org.uk/english/services/Pages/Freetranslationservice.aspx
Logainm	**Logainm (www.logainm.ie)** Since 2007, Fiontar has also been working together with the Placenames Branch of the Irish Department of Community, Rural and Gaeltacht Affairs to produce a comprehensive online database of Irish (including Northern Irish) place-names, building on the success of Focal. The public website was launched in October 2008 and is currently due to run until 2010. The site has both a concise and a detailed view that users can choose. The concise view gives both the Irish and English forms of place-names, the meaning and genitive forms. Sound files are also available for some place-names. The detailed view includes additional data, for example, where and when place-names were collected and different forms that may have been collected. Plans for the final phase include mapping tools, educational tools, additional information and an interactive system of gathering place-names information from the public. *See Fiontar for contact details.*
Maes-T	Maes-T is a terminology development management system. It is a web-based system through which widely dispersed teams of terminologists and language and subject experts can co-operate efficiently to develop and standardise terminology. It includes the following features: ▪ Full compliance with international standards ▪ Welsh lemmatiser ▪ Access and participation rights management The Maes-T system greatly reduces the amount of time required to work through a given amount of terminology. The way data is stored also facilitates the easy derivation of dictionaries, online or printed. www.maes-t.com

Name	Description
NCI & NEID	This is a jointly running project. One part consists of a corpus project, Nua-chorpas na hÉireann - New Corpus for Ireland (NCI). The second part is a new, corpus-based dictionary (Foclóir). The NCI itself consists of two distinct corpora: ▪ New Corpus of Hiberno English, approximately 25 million words ▪ Corpus of Modern Irish, approximately 30 million words, covering the period from 1883-date The Corpus of Modern Irish itself is based on the older National Corpus of Irish produced by the Linguistics Institute of Ireland (ITÉ). The National Irish Corpus itself was part of the European PAROLE project which produced harmonised core corpora for all European official languages between1996-1998. It initially contained 8.5 million words, to which another 15 million were added. As part of the NCI project, an extra 6 million words were added. Some of this material was obtained by appeals to Irish speakers for relevant material. Corpus text structure and data were encoded to be compatible with the specifications of the Text Encoding Initiative (TEI) and the Corpus Encoding Standard (CES). The corpus was also morphosyntactically annotated according to a common standard with language specific extensions. The work on the NCI was carried out by a contractor, Lexicography Masterclass (www.lexmasterclass.com), with the TCD Centre for Language and Communication Studies refining the parts of speech tagging processes. The NCI is not accessible online yet though it is possible to request access for research purposes via Foras na Gaeilge. The New English Irish Dictionary (NEID) is a bilingual dictionary project to address the need for a modern, comprehensive bilingual dictionary. Due for publication in 2012 in printed form and electronically, it will contain approximately 40,000 headwords. Phase 1 (planning and technical design was completed in 2006 and was also carried out by Lexicography Masterclass. This Phase included the creation of the NCI. Phase 2 (compilation and writing of the dictionary) commenced in 2008 and is carried out by the NEID team and Lexicography Masterclass. This phase will run until 2010/11. The NEID team consists of (lexicographical) editors from An Gúm, a corpus development officer, various technical experts (including Professor Scannell and people from TCD). Phase 3 (publication) will culminate in publication in 2012. This project also invites input from both speakers of Irish and Hiberno English to locate citations of words and expression that can be added to the database to ultimately present a fuller picture of usage. *See Foras na Gaeilge for contact details.*

Name	Description
Northern Ireland Place-Name Project	This project goes back as far as 1987, having grown out of the Ulster Place-Name Society which itself goes back to 1952. Over the years it has researched and published on a large number of place-names, including the Dictionary of Ulster Place-Names and volumes 1-8 of the Place-Names of Northern Ireland.[67] At the moment, people looking for Ulster place-names can approach the Ulster Place-Name Society (www.ulsterplacenames.org) enquiry service by post, phone and email only. Alternatively, the Logainm (q.v.) project contains a number of Ulster place-names that can be accessed online. QUB is currently working on making the data on Ulster place-names available online, including ancillary materials. This is done in collaboration and with the help of various bodies such as the Ordnance Survey for Northern Ireland and Foras na Gaeilge. There are also plans to link various place-names databases with a planned launch some time before the end of 2009. Given the similarity of the projects and the linguistic aspects, including the convention of using the Irish/Gaelic spelling of place-names in Irish/Gaelic texts, the possibility of using existing technology from either project; and linking the future database of Scottish place-names to the Irish projects should be vigorously investigated. Room 202, 7 University Square, Ollscoil na Banríona, Béal Feirste BT7 1NN Dr Kay Muhr, townlands@qub.ac.uk
Open Source projects (Welsh)	There are several Open Source software applications available for Welsh. In part these localisation projects are run by volunteers but some software or software utilities have been localised by non-volunteers. ▪ Operating System: Mandriva[68] Linux 2007 ▪ A localised version of OpenOffice 2.0 called Agored (www.agored.com). The project cost £320,000 and was run by the Mercator Centre. Financial support came from the Objective 1 European Regional Development Fund, the Welsh Assembly Government's Pathways to Prosperity Fund, the Bwrdd, S4C and Bangor University. 18.7% of the project costs were spent on translation itself. The project ran for two years and included supported trialling by SMEs, programming additional features (the Agored installation includes both the English and the Welsh interface and functionality so users can easily switch between languages) and a strong emphasis on publicity. ▪ Spellchecker: Gwirydd Sillafu Cymraeg (spell-checker) by the Canolfan Bedwyr for OpenOffice 3 ▪ Browsers: Firefox 3.5.1 (Windows, Linux, MacOSX) There are issues with running the last Welsh Thunderbird version with the latest version of Firefox. The current workaround is the

[67] 1992-2004, out of 30/40 planned volumes in total.

Name	Description
	installation of a Welsh language pack add-on for the English version of Thunderbird.
QUB	One of the projects at the Irish Department of Queen's University Belfast (QUB) is the Diploma/MA in Irish Translation Studies. Since 2003 QUB has offered a degree in Irish Translation Studies, available both full- and part-time (1 or 2 years respectively). Degrees in Irish translation studies are also available from other universities (Maynooth (online), NUI, DCU) but due to its location in Northern Ireland, the QUB course is particularly interesting to the Gaelic situation. The course was funded by EU money for the first 3 years. Students are assumed to be fluent and must pass a rigorous entrance exam, which also assumes basic computing skills. The uptake of the course has been reasonable, in particular with people just past retirement age who are looking to supplement their pension with translation work. The course used to teach a MT module but has ceased to do so as it was felt that too much time was spent explaining a particular software package that might not be the one required by a future employer, as there are dozens of systems. The general feeling regarding Irish spell- and grammar-checkers is that they are not good enough yet. Since the launch of the termbase Focal, specialist terminology is generally felt to be served well enough. In general, the department would be interested in collaborating or consulting a similar Gaelic project. As most Gaelic translators appear to work in translation part-time only, alongside a main occupation, any course developed must consider these limitations of time and movement. The use of e-Learning in some form may have to be considered. In this context it is worth noting that the current UHI Gaelic Plan (draft) appears to have identified Lews Castle College and Herriot Watt as its main partners for developing a course in translation. Collaboration with a provider whose main languages are Arabic and Mandarin Chinese (two major international languages) is perhaps questionable. Translators and interpreters working with Gaelic, an endangered and lesser-resourced language, have special needs and requirements which are not always comparable to those of professionals working with major languages. A partner with experience of providing training to speakers of a Celtic language (ideally Irish) is essential. Secondly, aiming for a "core module as part of UHI's Gaelic degree programmes at honours level" does not address the needs of existing Gaelic translators. Many Gaelic translators work on a part-time basis and have other existing work commitments. These circumstances could make it difficult for them to attend such a course.

[68] Formerly Mandrake

Name	Description
	In this view, the initial contacts made between Bòrd na Gàidhlig and Foras na Gaeilge in 2007 regarding a translators' accreditation scheme should also be pursued. Roinn na Gaeilge agus na Ceiltise, Ollscoil na Banríona, Béal Feirste BT7 1NN Dr Chris Dillon, c.dillon@qub.ac.uk; www.qub.ac.uk/irish-celtic/
Rannóg an Aistriúcháin	Although not initially conceived as a translation department, the Rannóg goes back as far as 1919. A translation service per se was set up in 1922 charged with the translation of laws both from and into Irish. In the first half of the 20th-century the Rannóg was also instrumental in the promotion of roman typefaces over the traditional typefaces and the simplification of spelling, culminating in the 1931 Spelling of Irish in Official Documents memorandum. The Rannóg's work soon also involved the coinage of neologisms to cope with the translation of documents from English into Irish. The following period saw switches from roman to Gaelic typefaces and an increasing amount of confusion regarding spelling conventions. Under de Valera, the Rannóg was charged with producing a simplified system of spelling that would bring the written language closer to the spoken language. This system was initially planned for use by the civil service but Litriú na Gaeilge: Lámhleabhar an Chaighdeáin Oifigiúil was eventually made available to the public as well in 1945. Especially via its use in the education system, the new orthography (known as the Caighdéan) increasingly supplanted the old spelling. This period also saw the production of guidelines on grammar and terminology by the Rannóg (such as Gramadach na Gaeilge agus Litriú na Gaeilge (1985). From 1972 onwards, the Rannóg also took on translation of European documents and the provision of simultaneous translation. However, the provision of digital documents did not take place until 2002. Computer-assisted translation tools were introduced in 2000/01 and, as a result, the Rannóg maintains its own large translation memories. The terminological work of the Rannóg and An Coiste has not always been ideally coordinated and this has led to some inconsistencies. A panel is currently (2007/09) being set up to review and expand the Caighdeán, with the aim of clarifying points not originally addressed (such as the treatment of long noun phrases) and to revisit various spelling conventions. An Coiste is also involved and has produced various discussion papers on known issues. Rannóg an Aistriúcháin, Oifig Thithe an Oireachtais, Teach Laighean, BÁC 2 Vivian Uíbh Eachach, vivian.uibheachach@oireachtas.ie; www.oireachtas.ie

Name	Description
SALT Cymru	The Canolfan have recently also succeeded in getting funding for running SALT Cymru. The aim of this group is to provide a networking space for academics, experts and people interested in Welsh language speech and language technology. SALT Cymru will: ▪ Provide information on developments in the wider field of SALT ▪ Distribute information on relevant events via the network, a regular newsletter and Murmur, the unit's blog ▪ Provide information, resources and help to researchers and developers It has to date also produced an extremely detailed report into the wider context of Welsh SALT entitled *SALTcymru, Project Closure Report - April 2008* (q.v.). www.saltcymru.org
Professer Scannell	Prof. Scannell is a professor of computer science at the University of Saint Louis, Missouri, a functionally fluent learner of Irish with a passive understanding of Gaelic. He also has a keen interest in the development of software tools for lesser-resourced and under-resourced languages. He has developed the following tools: ▪ GaelSpell*:* the first Irish spellchecker in 2000. The latest version (4.5) contains 33,062 headwords and 319,631 inflected forms. Produced as Open Source software, a modified version is now marketed by Cruinneog for Windows and is widely in use. ▪ An Gramadóir*:* the first Irish grammar checker in 2003 as Open Source software. This grammar checker is corpus-based (see An Crúbadán below). The 2003 version (0.1) operated on a 313,000 lexicon, 16 disambiguation rules and 146 grammar rules. By version 0.6 (2005), the engine was operation on 456 disambiguation rules and 1573 grammar rules. The Gramadóir currently deals best with 2 word grammatical issues but can handle strings up to 4 words long. Work continues to improve the grammatical "range". Since 2006, a Java version of An Gramadóir has been marketed by Cruinneog for Windows under the name Ceart. ▪ An Crúbadán*:* an automated web-crawler that quickly builds corpora from online sources for lesser-resourced languages.[69] This is currently being used to develop proofing tools for language such as Tagalog and Hiligaynon (Philippines), Igbọ (Nigeria), Akan (Ghana), etc. ▪ These corpora do not have gold-standard cores and are not tagged or annotated but instead rely on statistical analysis to extract the required information to overcome the lack of tagging, etc.

[69] This software is, however, not distributable.

Name	Description
	▪ Líonra Séimeantach na Gaeilge (Irish Semantic Network): a free 3D semantic network for Irish with over 77,000 individual word senses.[70] He is also collaborating on numerous other projects such as: ▪ Foclóir Nua Béarla-Gaeilge, which requires indexing and converting pre-Caighdeán texts into the modern spelling. (See also the Foclóir project) ▪ Managing/coordinating/participating in the localisation and maintenance projects for OpenOffice, Firefox, Thunderbird, Sunbird, KDE and Linux. ▪ ga2gd*:* Irish to Gaelic machine translation software. The first version was produced in 2005 with a lot of input from Caoimhín Ó Donnaíle at SMO. The current focus is on porting ("adapting") the current engine to the Apertium Open Source machine translation framework. Apertium[71] initially focussed on MT between closely related languages (Irish/Gaelic, Catalan/Spanish) but has begun to approach more distant language pairs such as Catalan/English, Basque/Spanish. Virtually all tools Professor Scannell has developed or participated in are Open Source and often this has led to follow on projects or products with added functionality being developed for the language in question. He is also very keen on supporting Gaelic to develop new resources and at the very least should be involved on a consultative basis in related Gaelic projects. Director of Computer Science, Department of Mathematics and Computer Science, Saint Louis University 220 N. Grand Blvd., Saint Louis, Missouri 63103-2007, USA Professor Scannell, kscanne@gmail.com; http://borel.slu.edu/index.html
Téacs	A need of young Irish speakers for predictive texting was identified and as a result, Foras na Gaeilge, in cooperation with Vodafone and the Tralee Institute of Technology, developed *Téacs*, a free predictive texting facility that works across a wide range of mobile phones.[72] Téacs contains some 25,000 Irish words commonly used phrases. The chief people involved were Carolan Lennon (Consumer Director at Vodafone Ireland), Ferdie Mac an Fhailigh (Chief Executive of Foras na Gaeilge) and Muiris Ó Laoire (muiris.olaoire@staff.ittralee.ie) and it can be downloaded and viewed at feedback@teacs.ie; http://teacs.ie/

[70] Semantic networks could be described as a "thesaurus on steroids".

[71] A spin-off from the Basque-led OpenTrad project which focussed on the language of Spain (Basque, Catalan, Galician and Spanish), see www.apertium.org.

[72] This may be related to the 2001 report *Ógshaothar - Staidéar ar ógsheirbhísí Gaeilge agus Gaeltachta agus ar riachtanais agus éilimh dhaoine óga*. We have been unable to determine the exact link but we would like to draw the attention of the Bòrd to the report, produced by Foras na Gaeilge and Údarás na Gaeltachta into the uptake and consumption of services by

Name	Description
Terminologia Batzordea & Euskalterm	The Terminology Council (Terminologia Batzordea) was set up by the Basque Government in 2002 to work towards bringing together and clarifying the various sources of terminology and to make them available online. In collaboration with all major stakeholders (HPS, the Government's Legal and Administrative Translation Service, UZEI, EHU, the Euskaltzaindia and the Education Department), the Council has to date worked through more than 60 dictionaries and other sources of terminology published between 1998 and 2005 and made the agreed terms available on Euskalterm (currently more than 187,000).[73] Search languages are Basque, Spanish, French and English (plus Latin for biological taxonyms). The Council does not develop terminology itself but it works towards standardisation, identifies gaps and priorities in terminology, consults with users, co-ordinates terminological work and disseminates the agreed terminology. This work is carried out in line with the Euskaltzaindia's recommendations and guidelines on terminology work. www.euskadi.net/euskalterm/indice_i.htm
Testun	Testun is a company that provides a number of Welsh language services, including subtitling, teletext and translation for S4C. The company is currently working on ways of speeding up the subtitling process and is interested in extending this technology to other languages.[74] Currently, live Welsh to English subtitling is done via a trained translator using (English) speech recognition software to produce English subtitles to live Welsh broadcasts. Current research is focussing on developing limited domain Welsh speech recognition which is then linked to a MT engine to produce an automated English translation. Testun Cyf, Tŷ Norfolk, 57-59 Heol Siarl, Caerdydd CF10 2GD post@testun.co.uk; www.testun.co.uk
To Bach	Draig Technologies, founded in 1999, is a provider of software services. The company has been involved in developing bilingual software guidelines for the Welsh Language Board and in the development of a number of Welsh Microsoft products. It has developed a software tool that facilitates the use of the circumflex on Welsh vowels (â ô û î ê ŵ ŷ) on computers running Windows. After installation, AltGr and any vowel letter will produce the equivalent vowel with a circumflex. Draig Technologies, Intec, Parc Menai, Bangor LL57 4FG Dr Richard Sheppard, info@draig.co.uk; www.draig.co.uk

young Irish speakers in Ireland and Northern Ireland, as an extremely interesting insight into the behaviour of young speakers of a Goidelic language. Foras na Gaeilge should be contacted for a copy of the report.

[73] Incluing departmental resources as well as published dictionaries and project termbases such as the Windows localisation termbase.

[74] This should, however, be seen within a wider discussion about the general desirability of more subtitling of Scottish Gaelic programs.

Name	Description
Tobar na Gaedhilge	Tobar (currently version 1.4) is a private corpus project started by Ciarán Ó Duibhín in the 1990s. It contains 20th-century material for Ulster, Munster and Connacht Irish and Gaelic. It currently contains approximately 3 million Irish words and 100,000 Gaelic words. The Gaelic material consists mostly of 4 hand-typed books - *Trì Dealbhan Cluich* by Alasdair Caimbeul (1990), *Companach na Cloinne* by Iain MacPhàidein (1912), *Am Measg nam Bodach* (1938) and *Seanchaidh na Tràghad* by Iain MacCormaig (1911). Some of the Irish material has been incorporated into *Corpas na Gaeilge 1600-1882* by the Royal Irish Academy. The engine is capable of dealing with lenition but does not contain a lemmatiser (except for English and French, as the corpus contains English/French translations of some of the material). None of the material is tagged. It currently runs on Windows only (on MacOSX only via a Windows emulator). Ciarán Ó Duibhín would be willing to share the Gaelic data with other Gaelic corpus projects. The University of Ulster intends to develop a Corpus of Written and Spoken Ulster Irish, based on TnaG. The aim is to add another 5 million words of Ulster Irish to the existing 3 million work corpus. 165 Andersonstown Rd, Béal Feirste BT11 9EA Ciarán Ó Duibhín, ciaran@oduibhin.freeserve.co.uk; www.smo.uhi.ac.uk/~oduibhin/tobar/
Traslán	Traslán was set up in 2004 as a translation and web design services provider with a strong emphasis on technology and training. They have worked together with a large number of eminent Irish institutions such as Trinity College Dublin (TCD), Fiontar from Dublin City University (DCU), Foras na Gaeilge and the Department of Community, Rural and Gaeltacht Affairs (CRAGA). As part of their technology base they have developed Irish machine translation (MT) software, are involved in training translators, in developing Irish translation memories (TM) and they maintain a large online terminology database. Overall, Traslán's assessment of the situation regarding technological tools and terminology is fairly positive. Many useful tools are already at the disposal of Irish translators, so the main challenges remaining in their view are the provision of better training, dissemination of terminology and technology and developing the corpora. **Irish MT** Traslán's MT software was developed over a period of 4 years by the company. It is a hybrid MT example and statistics based system, thus reliant on large bilingual corpora. A rules-based approach was deemed not to be workable enough during early development. A lot of the preliminary work around MT, however, was carried out over a 10-year period under the auspices of Andy Way,

Name	Description
	associate professor in computing at DCU, editor of the Machine Translation Journal, member of the National Centre for Language Technology (NCLT)[75] and the European Association of Machine Translation (EAMT).[76] Also involved was Carl Vogel of TCD at the Centre for Computing and Language Studies and Josef Van Genabith at DCU and a number of other students. The initial corpus used was based to some extent on the former (monolingual) National Corpus of Irish developed by the Linguistics Institute of Ireland (ITÉ). Traslán's current in-house bilingual corpus contains some 32 million tokens. The current version of the MT software can handle roughly 1000 words per minute. As with machine translation in general, it is more adept at handling non-literary texts such as software projects, public sector documents and legal documents. It is less successful with literary texts such as fiction. Manual proofreading is still required. Traslán's MT software is currently only used as an in-house tool. However, it has been used for other language pairs (Mandarin, German, French, Arabic, etc) with reasonable success depending on the size of the bilingual corpus provided. However, Traslán has also indicated that it would be delighted to co-operate with the Gaelic community to explore the potential of using it for Gaelic MT, provided bilingual corpora or TMs can be made available. **TM Development** Traslán is also involved in a project with Foras na Gaeilge and CRAGA to produce and distribute TMs and to provide training in using them. At this stage only a generic TM will be used, with an emphasis on public service related language due to the increased demands as a result of the Official Languages Act (2003) and Irish becoming an official EU working language (2007). The TM will be provided in .tmx file format which will ensure good cross-platform support. Traslán will also provide OmegaT, a free computer-assisted translation (CAT) programme (running on Windows, Mac OS X and Linux) as part of the package to encourage better uptake of CAT/TM, which is currently described as extremely low amongst Irish translators. This will be in conjunction with Foras na Gaeilge's own translation technology awareness campaign. However, the TM contents will also be available as a tabular text document for translators wishing to use it as a reference wordlist. Traslán, 31 Garrán an Mhuilinn, Coillín na Carraige, Na Clocha Liatha, Co. Chill Mhantáin, Éire Donncha Ó Cróinín, docroinin@indigo.ie; www.traslan.ie
UZEI	UZEI (The Basque Educational Centre for University Services) was founded in 1977 as a NfP organisation with the stated aim of developing Basque terminology to (ultimately) enable Basques to use the language to discuss any given topic. From 1987

[75] www.nclt.dcu.ie

[76] www.eamt.org

Name	Description
	onwards, it has been recognised by the Government of the Autonomous Community as a trustee for research into linguistic planning and from 1989 as a public interest body. It has consulted, cooperated and developed (and continues to do so) a large number of projects such as: ▪ Euskaltzaindia's planned Hiztegi Orokorra ▪ Corpora, in particular the design and lemmatisation of the corpus of 20th-century Basque for the Euskaltzaindia (q.v.) ▪ MultiMeteo, the weather forecast system (see EHU) ▪ euLex, UZEI's lexical database of Basque UZEI is a member of various sector organisations and regularly organises conferences, workshops and training events (including the tertiary level) on terminology development, corpus design and its other specialisms. **Euskalterm (www.euskadi.net/euskalterm/indice_i.htm)** UZEI sat up Euskalterm in 1986 as a digital termbase of technical terms based on UZEI's large number of technical dictionaries published from the 1970s onwards. These include highly specific dictionaries on topics such as museology, the digestive system and handball. Control of Euskalterm was passed on to the Government's Language Policy Department in 2001 with UZEI remaining a collaborative partner in the project. UZEI remains the main body charged with creating new terminology. **TEIS** TEIS is a Terminology Implementation Information System developed by UZEI, based on a methodology developed Jean Quirion at the Université du Québec en Outaouais.[77] It measures the uptake of new terminology in specific domains. For example, using this technology UZEI has been able to determine that the uptake of terminology sanctioned by the Euskaltzaindia within formal domains, based on a 500,000 word corpus collected for the purpose, is above 94% overall and specifically 95.4% within the education system, 94.1% within the administration, 93.9% within the media sector and private companies. In the current, second phase of the project, a web-based corpus will be used to evaluate the wider uptake. UZEI, Aldapeta 20, 20009 Donostia, Gipuzkoa, Spain uzei@uzei.com; www.uzei.com

[77] www.uqo.ca

Name	Description
Vifax	Vifax is a language teaching/learning tool developed and maintained by the National University of Ireland,Maynooth. Each week, it takes a news broadcast from the Irish language channel TG4 and produces learning materials and a transcription for 3 different learner levels. These, including a recording, are made available on the university's website free of charge for learners. Originally these were produced from videotaped episodes of news broadcasts but since the arrival of digital recording technology, the process has become much easier and quicker. Ionad na dTeangacha, Ollscoil na hÉireann Má Nuad, Má Nuad Pádraig Mac Gabhann, padraig.macgabhann@nuim.ie; www.nuim.ie/language/vifax/index.shtml
WISPR	The WISPR (Welsh and Irish Speech Processing Resources) project grew out of the needs of the Welsh speaking disabled community. A project to produce Welsh and Irish TTS software (see *Abair*) was set up using European Interreg IIIa (Wales and East of Ireland) funding in the Interreg 2000-2006[78] period. Work was carried out in collaboration with Trinity College Dublin, with support from Dublin City University, University College Dublin and the ITÉ. The project has to date produced 3 Welsh diphone voices (the initial voice being of lower quality) between 2004 and 2006. The voices were built using the freely available Festival engine and are freely available on the website (including technical documentation). There are plans to update the system by moving from the current diphone system to a unit selection or hybrid model. www.e-gymraeg.org/wispr/index_en.htm

Links to other relevant organisations:
- Århus Centre for Lexicography (www.asb.dk/article.aspx?pid=893)
- European Center of Excellence in Speech Synthesis (www.ecess.eu)
- LangTech (www.lang-tech.org)
- Language Technology World (www.lt-world.org)

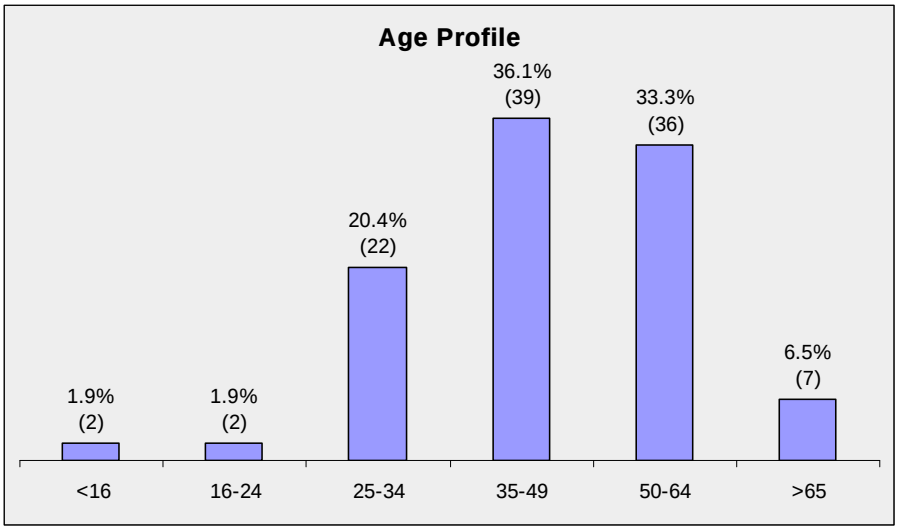
[78] The current period is 2007-2013, see http://ec.europa.eu/regional_policy/funds/2007/index_en.htm.

Appendix 3

Online Survey Results

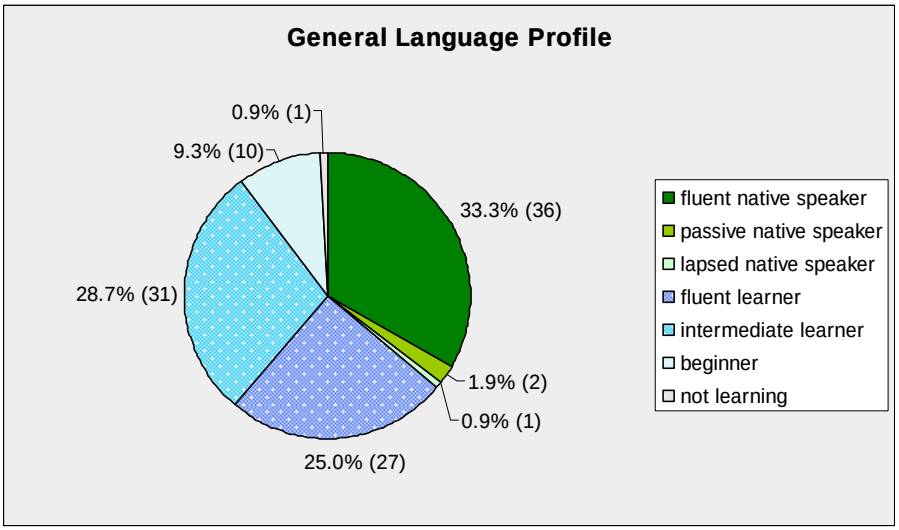
Question 1 - Age Profile of Respondents

Out of 121 total responses, 108 had usable data past the profiling page. Although most responses were from people aged 25 or over, there were also 4 responses from people aged 24 or younger.



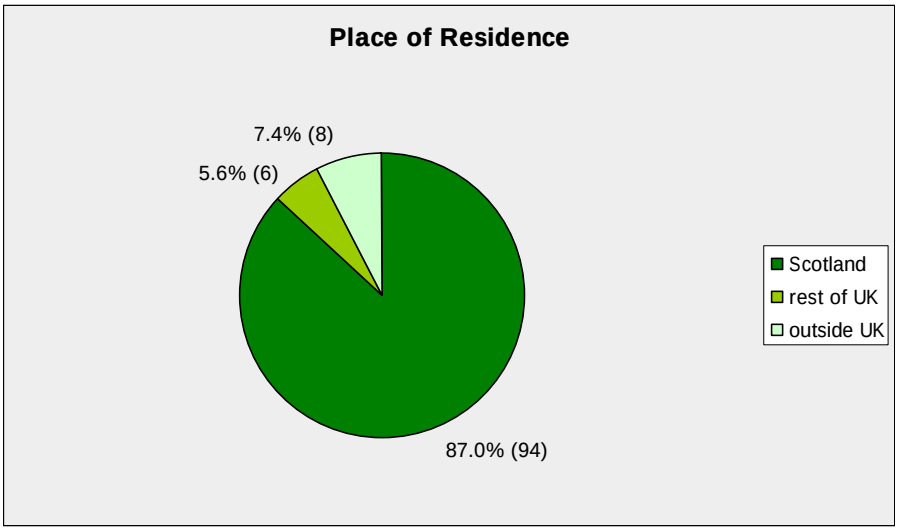
Question 2 - General Language Profile

Over a third (36.1%) of respondents were native speakers of some type, 63% learners.



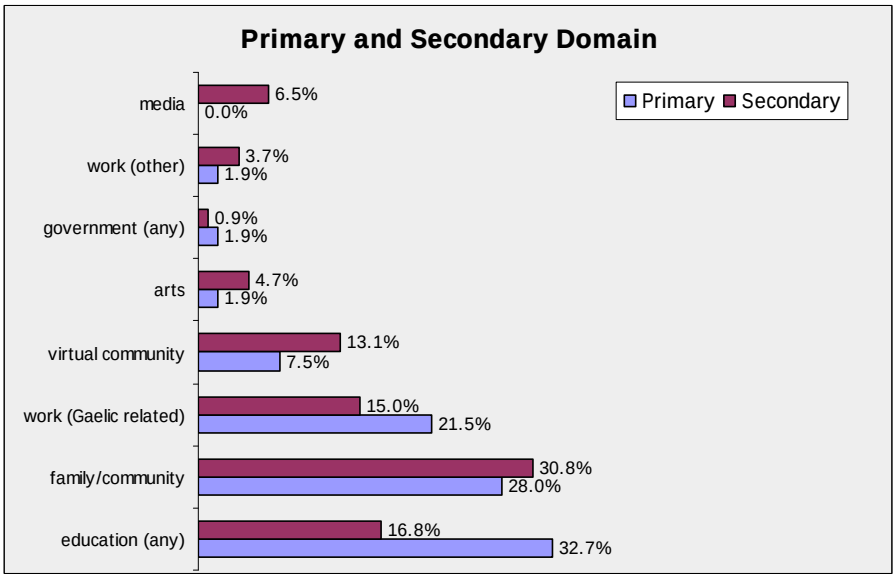
Question 3 - Place of Residence

The vast majority (87%) was resident in Scotland, with 7.4% living outside the UK.



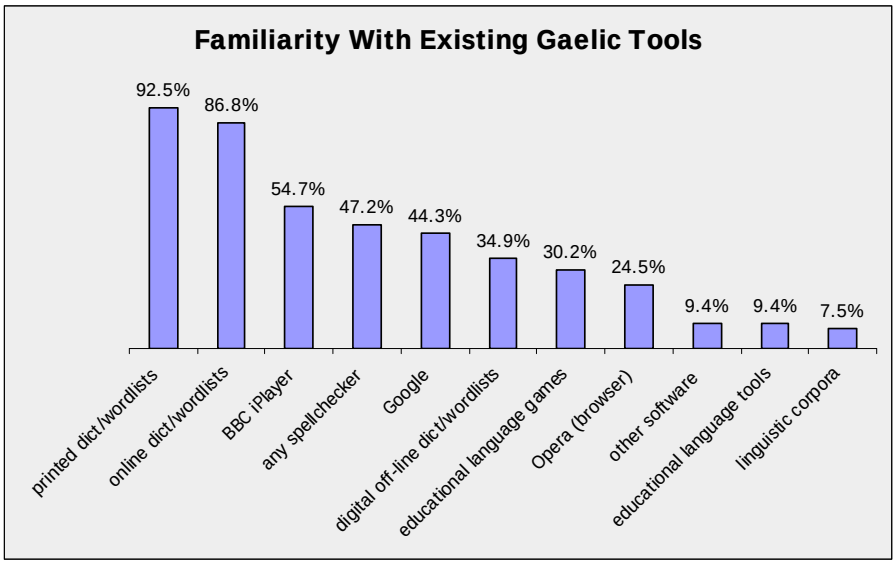
Question 4 - Primary and Secondary Domain of Language Use

Respondents were also asked to state the primary and secondary domain of their language usage. For the majority (32.7%) this was the education sector (primary, secondary, tertiary or adult), followed by the family/community (28%). The fact that Gaelic use in the virtual domain exceeds that of Gaelic use in the arts, government and media is notable but it should be borne in mind that this was a web-based survey and thus does not reflect language use amongst non-web-users.



Question 5 - Familiarity with Existing Tools

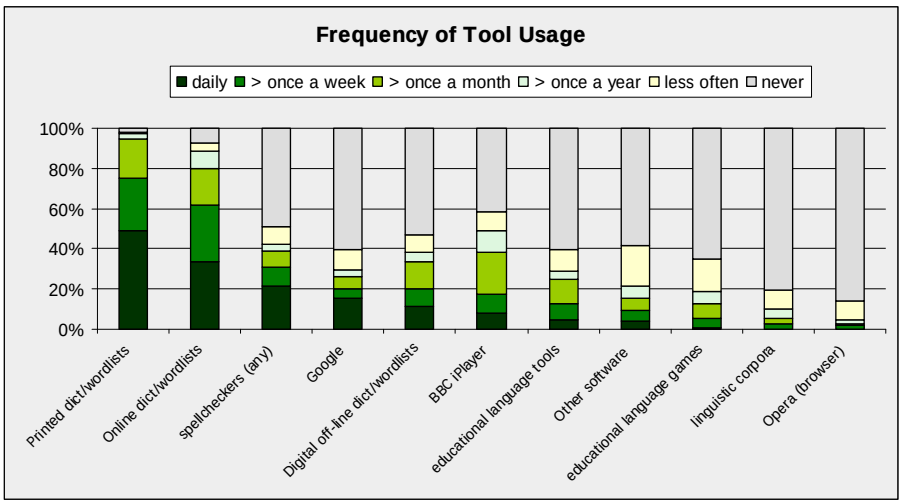
Familiarity with existing tools varied. It was highest with printed and online terminology resource (though low with off-line digital resources such as Roy Wentworth's Wester Ross Gaelic Dictionary). Other resources (either online or available online) were known to about half of the respondents. The low figure for linguistic corpora is likely linked to the fact that they are specialised tools and, in the case of Gaelic, extremely small and thus not overly useful even to specialists.



Question 6 - Frequency of Gaelic Tool Usage

Usage of tools respondents were familiar with showed:

- A high uptake of terminology resources (75% daily or weekly usage of printed resources, 62% daily/weekly usage of online resources). These categories also had the lowest (1.6%) incidence of "never".
- Comparing the usage of Google's Gaelic interface, BBC iPlayer and Opera could support the view expressed by several respondents that limited functionality in the Gaelic version of a tool restricts usage. The Gaelic version of Opera is several years out of date (last version 4.02, 2000); of the 20 functionalities immediately available on Google.co.uk, only 5 (Search, Images, Groups, Directory and Login) are available in Gaelic; BBC iPlayer offers the same level of functionality in all versions.
- Corpora usage was extremely low; however, all existing Gaelic corpora are currently too small to be useful to researchers.



Question 7 - Reasons for Non-use of Existing Tools

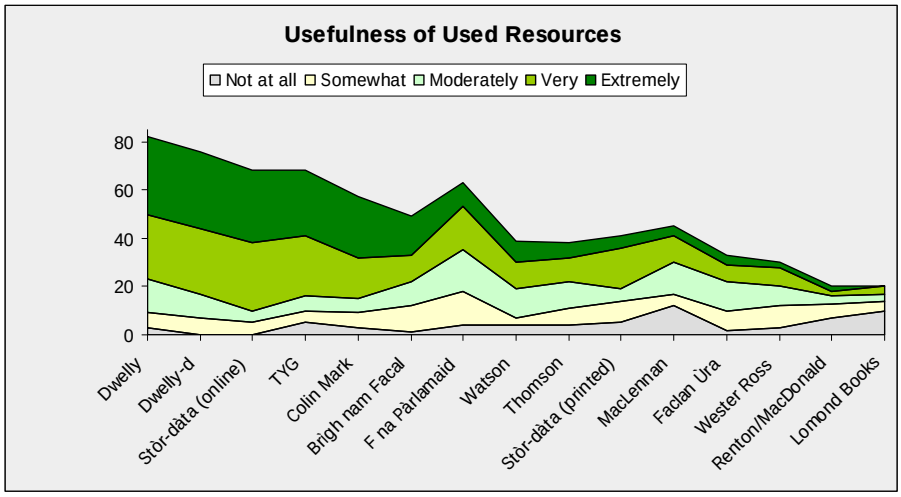
There were 98 open responses to the question of why respondents never used a resource available in Gaelic. The most common response (65%) was that they simply were not aware of their existence. This includes resources that the team expected to be well-known such as the *Dearbhair* spell-checker. Various respondents expressed their gratitude to the research team for drawing their attention to the existence of certain resources. Ignoring the "not required in my work/life" response, other common responses were:

- Too complicated to install/set up.
- Technical problems operating a resource.
- Lack of trust in the resource (spell-checkers in particular were mentioned).
- Lack of functionality or out of date compared to the same/comparable tool in English (in particular the Gaelic web-browser and OpenOffice).

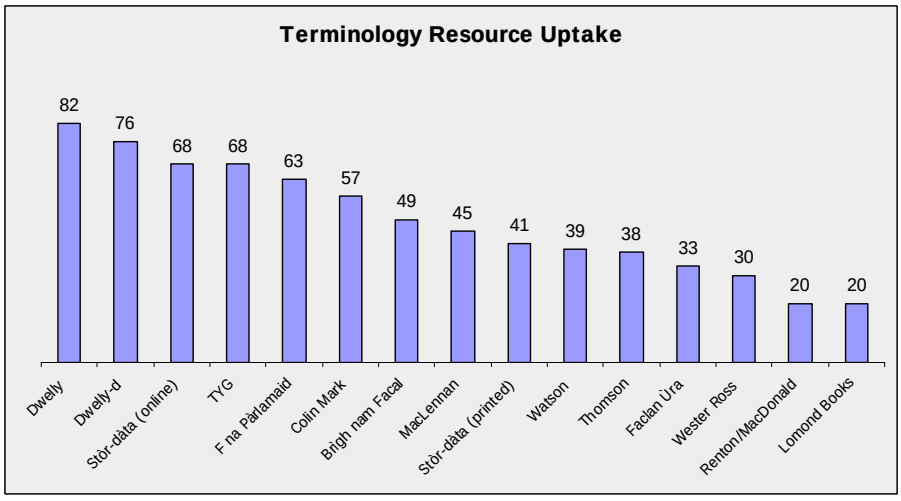
Question 8 - Evaluation of Used Tools

Respondents were asked to rank their perceived usefulness of those existing terminology resources they used at least several times a month.

- In spite of its age, Dwelly is perceived as the most useful resource (digital or printed). This is closely followed by the online Stòr-dàta and Colin Mark's Gaelic-English dictionary.
- In spite its small size, the TYG[79] dictionary, the only sizeable modern bidirectional dictionary, is ranked high.
- Faclan Ùra (Gaelic - English wordlist), currently the only sizeable source of school terminology, is only used by 31.4% at least several times a month and only a third of those (10% of overall respondents) consider it to be extremely or very useful.



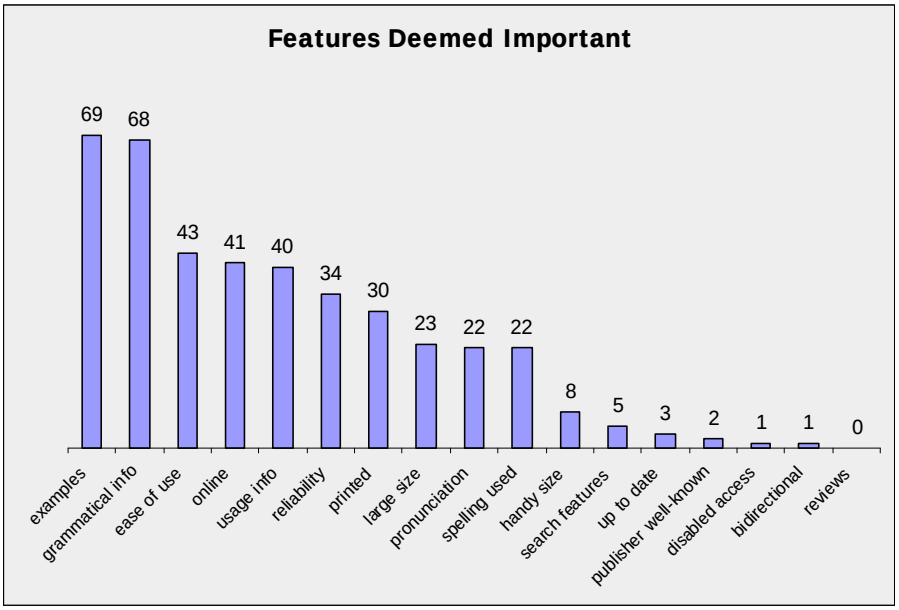
The analysis of the uptake of terminology resources based on the number of evaluation responses to each individual resource suggests that some (significant) resources such as *Faclan Ùra* are not well known or are not used frequently. The fact that it is nor readily available may also be a factor in its low usage.



[79] Robertson, B. and MacDonald, I. Teach Yourself Gaelic Dictionary, Teach Yourself Books 2004

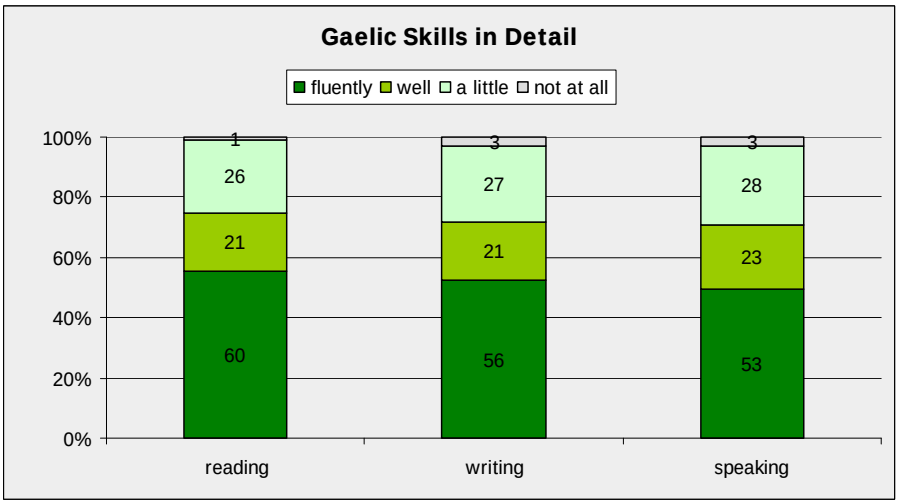
Question 9 - Dictionary Features Deemed Useful in General

In response to the question of what four features users considered are most important in a dictionary; examples and grammatical information were deemed the most important features by far. The next group of features considered most important by respondents were ease of use, availability online and information on usage (such as marking of neologisms, regional words, etc).

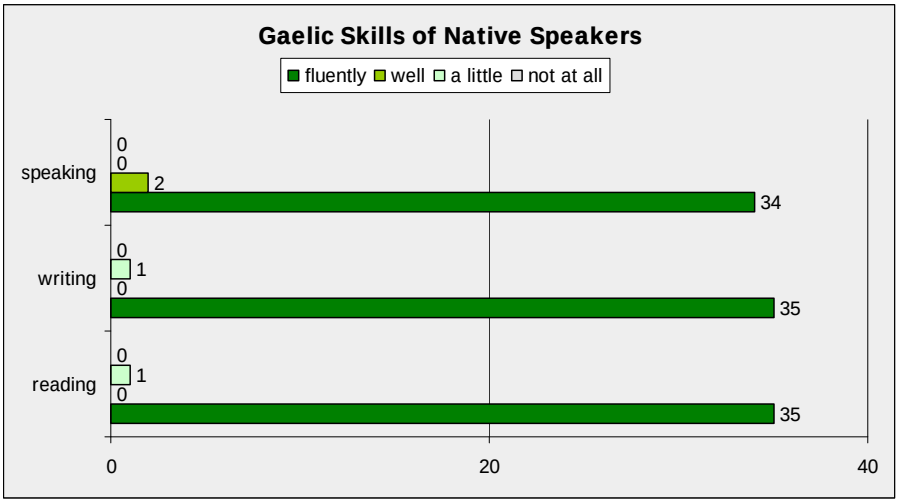


Question 10 - Speaking, Reading and Writing Skills

The speaking, writing and reading skills of respondents indicated a relatively high degree of literacy amongst respondents.



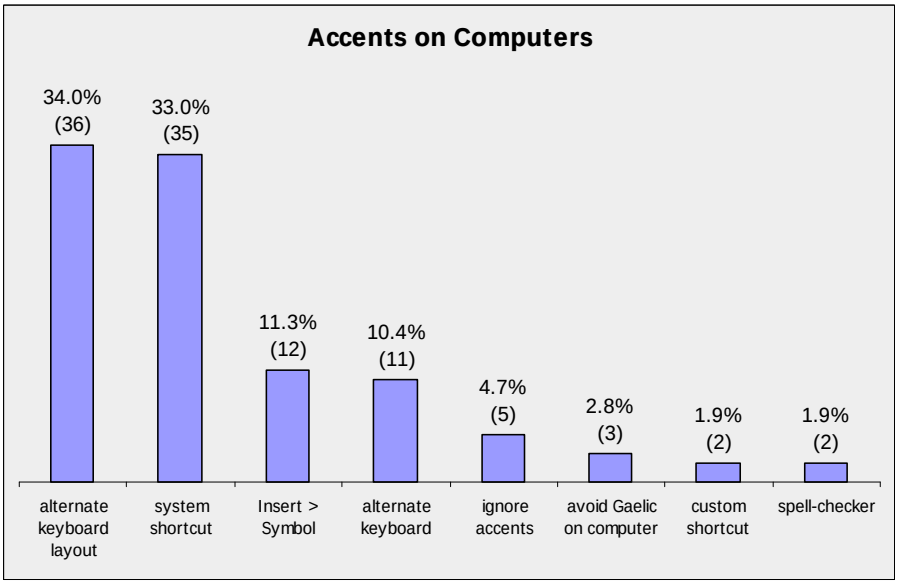
Amongst native speakers, the distribution was particularly high, although this is likely due to the high percentage of Gaelic professionals in the sample (including 19 translators alone).



Question 11 - Approach to Accented Characters on Computers

The question on how people deal with the accented letters when using computers revealed that of 106 respondents:

- 34% of respondents use a system keyboard layout (Irish, British extended, International, etc) that allows the insertion of accented letters via combining characters in any application. This figure includes the pre-defined keyboard shortcuts available to Mac users.
- 33% use a system shortcut (such as ALT 0224).
- 4.7% completely ignore accents and 2.8% of respondents actually avoid using Gaelic on computers because they cannot handle the accented letters easily.
- Several respondents stated in the open response section that their use of accented characters depends on the domain. A frequent response was that while people would use accented characters in applications like Word, they would ignore them in emails.



Question 12 - Satisfaction with Approach Used

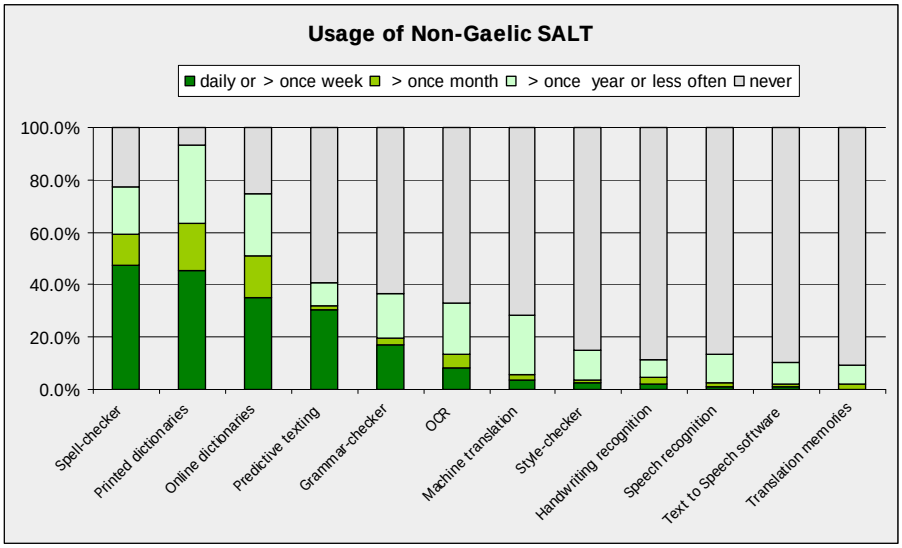
Overall, 60.2% indicated they were happy with their set-up. Of these, the majority (63%) indicated they were using an efficient way of entering characters (alternate keyboard layout or alternate physical keyboard layout).

The results also indicate that half of all respondents (51.8%) use overly complicated or time consuming methods (system shortcuts, Insert > Symbol (plus copying & pasting), spell-checker) or ignore them altogether.

Question 13 - Use of Non-Gaelic SALT

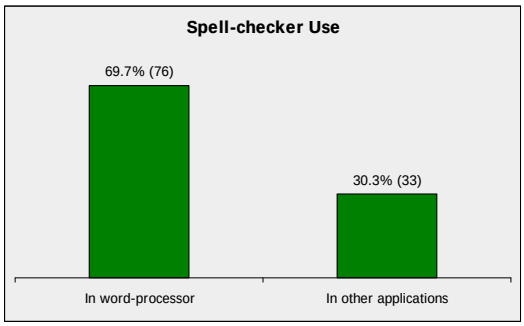
For comparative purposes, questions were also posed on the respondents general usage of non-Gaelic SALT outwith the narrow range of available resources in Gaelic.

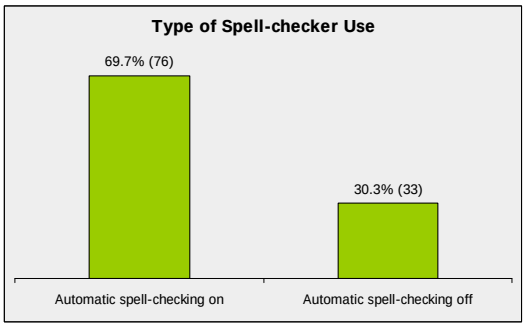
Proofing tools (64.7%), online or printed dictionaries (80.2%) and predictive texting (30.2%) were the tools most commonly used by respondents daily or several times per week.



Question 14 - Approach to Spell-checkers

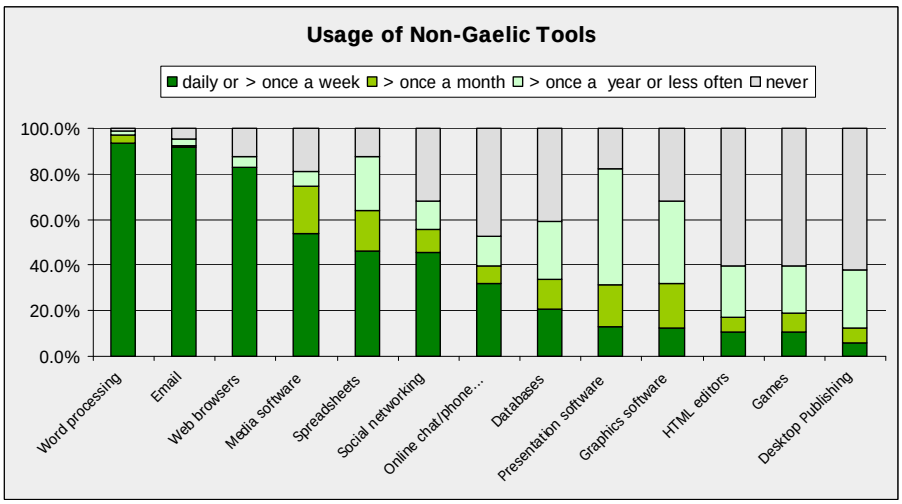
Of all spell-checker users, 69.7% used them in word-processors only and 73% had automatic spell-checking switched on. However, this includes a number of responses indicating that usage and setting varied depending on work/home use of a computer.





Question 15 - Frequency of Non-Gaelic Software Usage

The tools used by respondents daily or several times a week were word-processing (93.4), email (91.5%) and web-browsers (83%). The next cluster consisted of media software (53.8%), spreadsheets (46.2%) and social networking (45.3%).

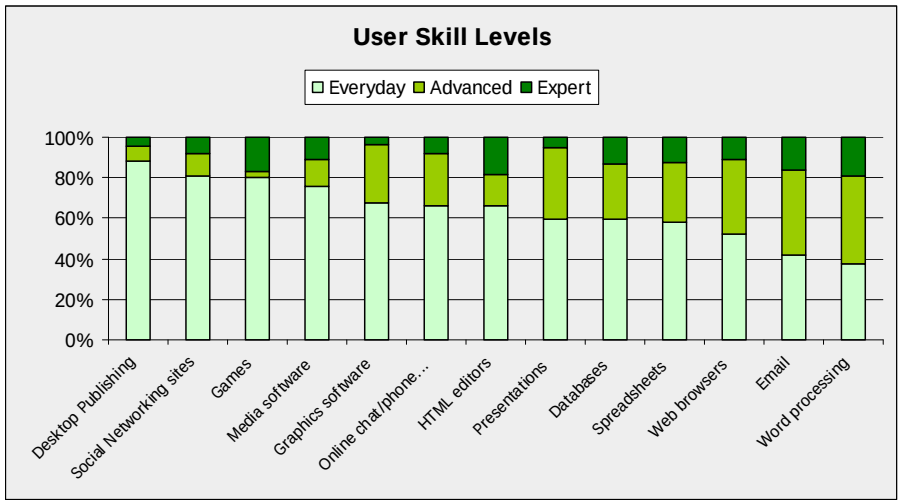


This indicates that the majority of respondents use a relatively narrow range of tools (word processing, email and web-browsers) most frequently. Some technologies were used very rarely only although in the case of games, this may well be due to the relatively low number of young respondents.

It suggests strongly that within this target audience, the emphasis of development needs to be on commonly used tools such as word processing, email and web-browsing software rather than more infrequently used tools.

Question 16 - User Skill Level

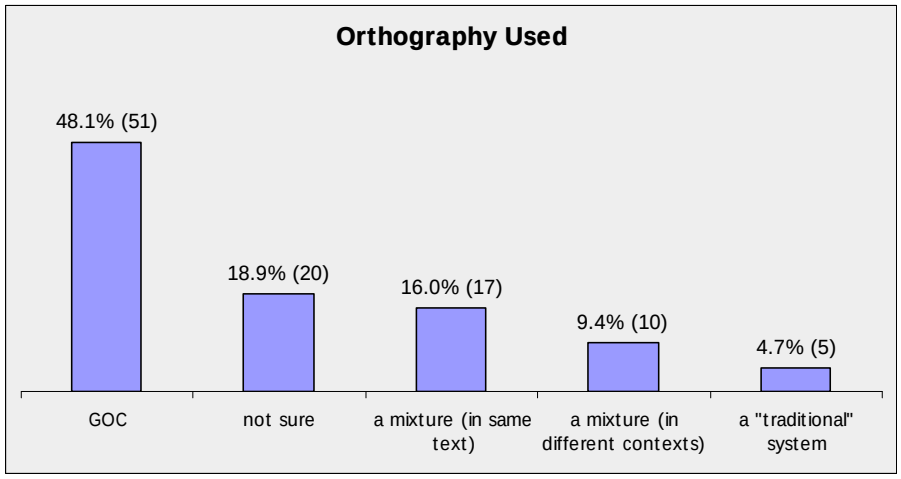
The majority of respondents described their skills at using different types of technology as everyday in the vast majority of cases. The only categories where a majority of people described their skills as advanced or expert are email (58.1%) and word processing (62.4%).



This indicates that any type of new Gaelic SALT that is to succeed must, from the outset, be designed to be easy to install and handle. Early testing by everyday users should be considered compulsory.

Question 17 - Type of Orthography Used

Just under half (48.1%) of respondents stated they used GOC. About a quarter (24.4%) were mixing systems, either within the same text or in different domains (e.g. GOC at work, another system privately). A significant number (18.9%) were simply unsure.



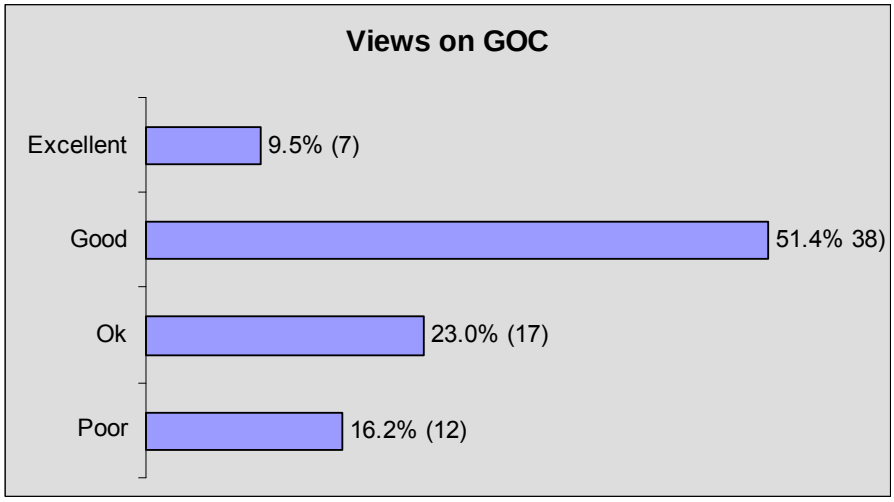
GOC usage amongst translators was high (94.7%). Excluding translators, only 44% consistently use GOC and 23% simply did not know what spelling they were using. As the figure excluding translators includes a number of GME teachers, usage rates amongst the general population are likely to be even lower.

Question 18 - Experience of GOC

The 74 respondents who were aware of the existence of GOC and the "traditional" systems were asked to rank their experience of GOC into one of four categories:

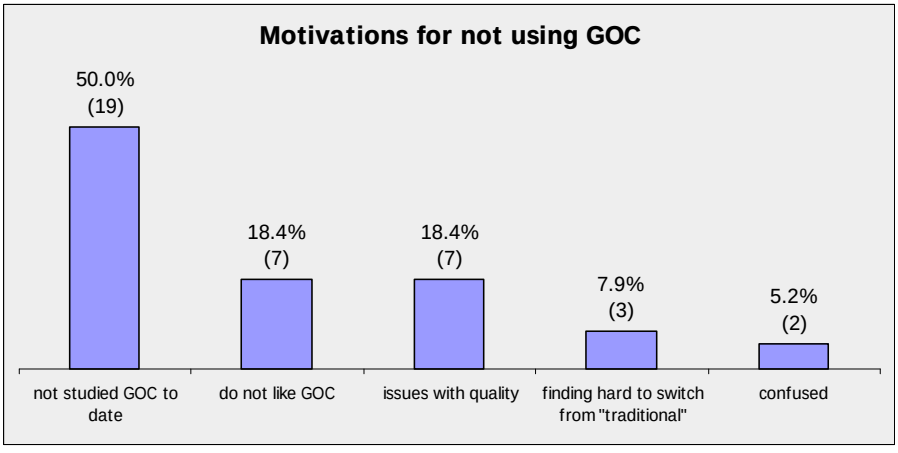
- Excellent (all issues are addressed and explained properly).
- Good (many issues are addressed and explained, needs a bit more work).
- Ok (some issues are addressed and explained, needs more work).
- Poor (not comprehensive enough, not well explained, needs a lot more work).

9.5% rated it excellent, 51.4% good and the remaining 39% OK or poor. This indicates that the vast majority of GOC users perceive the framework and its explanation as being less than perfect.



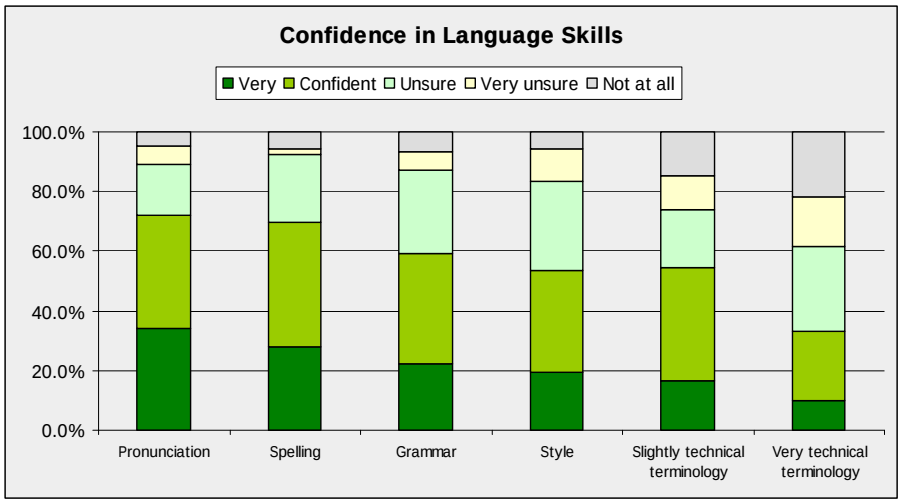
Question 19 - Reasons for Non-use of GOC

Of the non-users of GOC, 50% stated they had not had the opportunity to familiarise themselves with it to date; the other two main reasons were a dislike of GOC and qualitative issues.



Question 20 - Language Skill Confidence

Asked about their confidence in handling different domains of the language, the issues respondents felt least confident with was technical terminology, style and grammar.



Question 21 - Open Question on Standardisation

Respondents were also invited to express their views on any aspect of standardisation in Gaelic that they would like to comment upon. This is clearly an emotional issue for many people and the question attracted many open responses, some of them of considerable length. Of the 52 open responses, the following points were raised several times:

- The need for accepting a standard and stability of the standard (13).
- The need to further develop and refine GOC and to eliminate errors and exceptions (8).
- A general lack of guidance on a variety of issues such as advanced grammatical issues and spelling (7).
- The need for an independent "Academy" to work on standardisation (spelling, grammar, terminology) based on research and staffed by experts. *An Seotal*, a terminology initiative run by Stòrlann, was welcomed in principle but criticised more than once for a lack of transparency and lack of involvement of experts in terminology/lexicography (8).
- The need for more transparency in standardisation, better communication/ consultation and a more "gentle" approach to introducing the general public to it (6).
- A need to ensure that a standard is clearly delineated from "everyday usage" of the language and to ensure it does not restrict the traditional richness of the language and its dialects.
- Requests for re-visiting the acute accent (5).
- An urgent need to address the confusion of technical terminology, including place-names, surnames and biological taxonomy (5).
- Evaluating the rationale for the creation of new ghost words and ad-hoc renditions of loanwords (3).[80]

According to one respondent's experience in the education sector, children in GME are not exposed to pre-GOC systems enough to enable literacy in pre-GOC publications.

Other responses indicated a greater need to focus not only on terminology but also good idiom, better pronunciation, the distinction of registers and overall, more and better guidance and information.

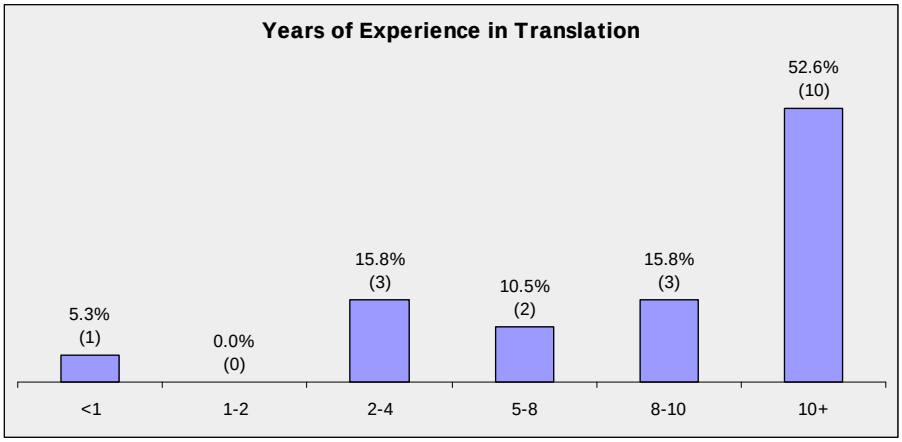
[80] Such as *dhad, led*, etc.

Question 22 & 23 - Gaelic Translators & Qualifications

19 respondents (18.4% of the total) stated they were either part- or full-time Gaelic translators translating more than 1,000 words per week on average. Of these, none held any professional qualification in translation.

Question 24 - Work Experience

10 of the 19 (52.6%) have worked in Gaelic translation for more than 10 years.



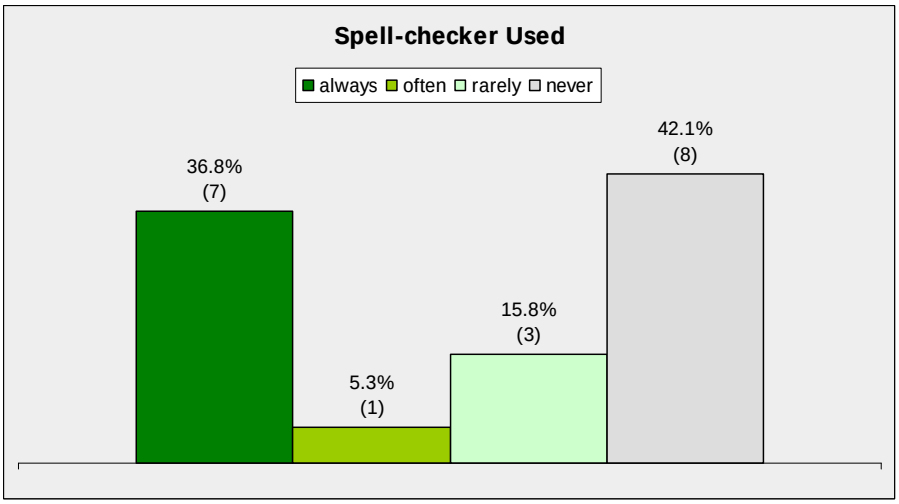
Bearing in mind the small sample size, the fact that there was only 1 translator in the <1 to 1-2 bracket may indicate a problem with young Gaelic speakers being attracted to the industry and warrants further investigation.

Question 25 - Type of Employment in the Translation Industry

More than half (52.9%; 9 respondents) stated they worked as an employee for an organisation that required Gaelic translation work of them. 41.2% (7 respondents) worked as part-time freelance Gaelic translators and 5.9% (1 respondent) as full-time freelance Gaelic translators.

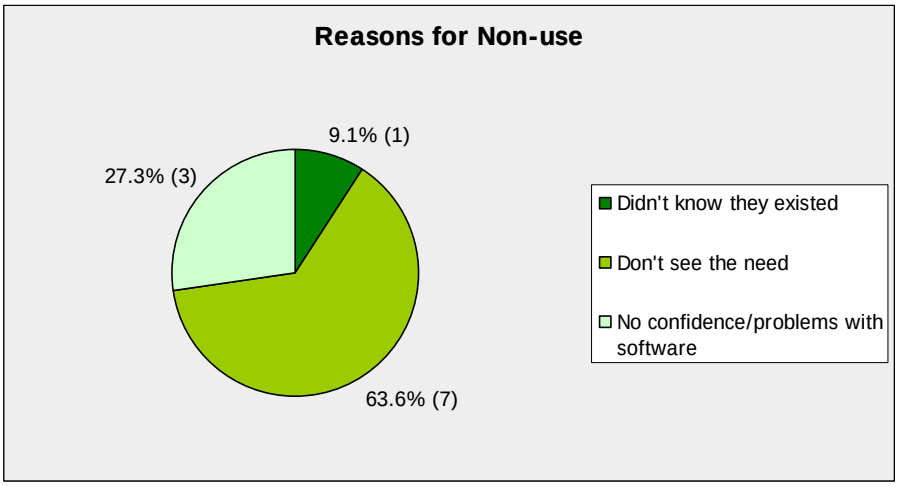
Question 26 - Use of Gaelic Spell-checker in Translation

When doing Gaelic translation work, 11 of the 19 (59.9%) stated they never or rarely use a spell-checker. This is a curious result and points towards a failure on behalf of clients requiring Gaelic translations to require work to be spell-checked. In non-Gaelic translation work this is normally a standard requirement and part of the terms of work.



Questions 27 - Reasons for Non-use of Gaelic Spell-checker

The translators not using spell-checkers stated they did not see the need for one and that they trusted their own abilities to spot errors. The second main reason was a lack of confidence in spell-checkers and problems with the software. This included complaints about the lack of updates and the inability of the software to "learn".



Question 28 - Relative Importance of Terminology Resources

The translators were also asked to rank the existing tools in terms of importance to them when doing English to Gaelic translation work. It revealed the following:

- The most highly ranked resources by the most translators were the online Stòr-dàta and Dwelly-d.
- The printed Faclair na Pàrlamaid was ranked higher than the online version.
- Using Google to locate a term was surprisingly common.
- MacLennan's dictionary received the highest number of bad rankings.
- A number of translators are using Irish dictionary resources.

Question 29 & 30 - Search Behaviour

The majority (57.9%) will consult a few sources in order to make sure they have the correct term. A minority (21.7% in each case) will either go with their first source or search for as long as necessary in order to locate a term.

Question 31 - 38 Use of TM Software

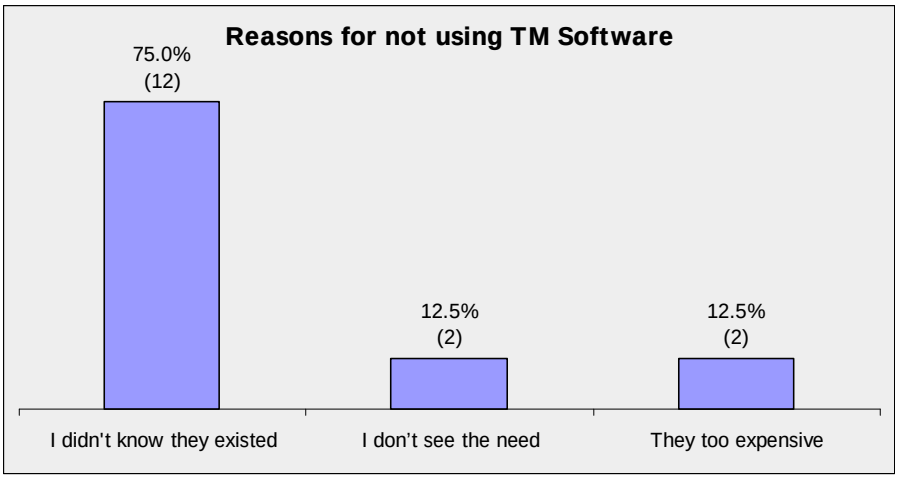
Only 1 of the 19 translators stated they used translation memory software; namely the proprietary Trados and the Open Source Poedit. TM software was judged "somewhat useful" and the reason for using it was given as clients requiring its use. Other statements were:

- That Trados was too expensive.
- Only a few key features were used by the translator.
- They never received training and would not want training.

The sample for question 31-38 was extremely small (19 responses). However, the recent list of Gaelic translators compiled for Bòrd na Gàidhlig lists a total of 42 Gaelic translators, so the sample may actually represent a significant percentage of all Gaelic freelance translators. It is also broadly in line with the results of similar research carried out into the uptake and opinions of TM software within the UK freelance translation market.[81]

Question 39 - Reasons for Non-use of TM Software

Of the remaining translators who do not use TM software, 75% had simply no idea they existed or what they did. One commented the lack of existing TMs and agreement of which software was used most in Scotland was making them hesitant over which to use.



[81] See *Translation Memory Survey 2006*, Imperial College London in the Attached Files.

Question 40 - Wish list of Tools

In an open question, translators were invited to state which tools they missed most and would be most useful to them (numbers in brackets indicate number of requests):

- An online dictionary/improved Stòr-dàta which includes grammatical information, examples and idiomatic usage (7).
- An interactive, standardised online terminology database (including place-names) (6).
- Affordable translation memory software (4).
- An online thesaurus (3).
- An authoritative and comprehensive grammatical description, including advanced topics such as treatment of long noun phrases (3).
- Guidance for translators and a resource site (2).
- A solution to the accents issue (2).

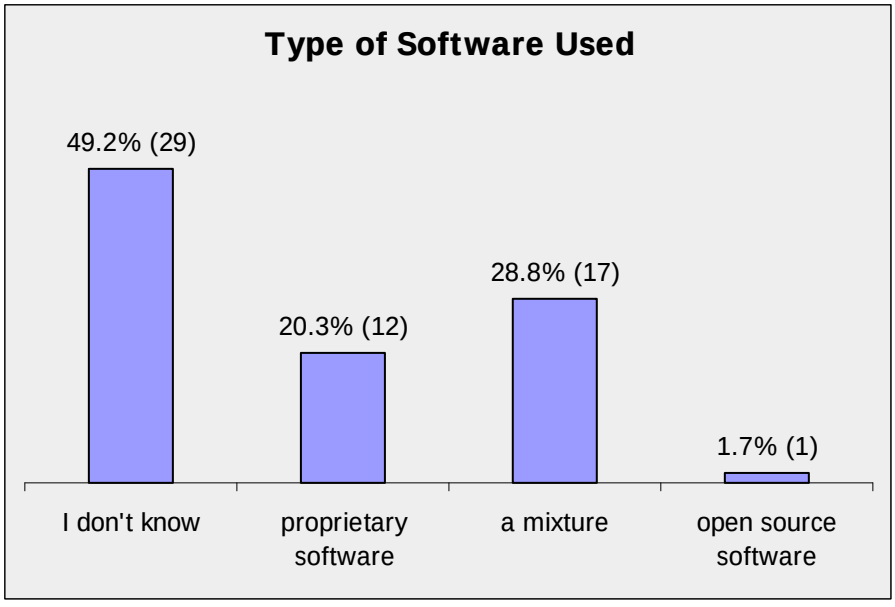
Other requests were made for a translator's online forum, a term extractor, downloadable TMs, guidance on acronyms, a "bigger and better" spell-checker.

Question 41 - Gaelic at the Workplace

59 respondents (57.3%) stated they worked for an organisation where Gaelic was an integral part of everyday work life.

Question 42 - Software at the Workplace

In terms of software used within the workplace, almost half were unaware of the type of software (proprietary vs Open Source) used. A fifth (20.3%) stated the software used was proprietary and a third (30.5%) stated a mixture or only Open Source software was used at work.



Question 43 - Potential Perceived Impact of Gaelic SALT

Asked to rank (Gaelic) tools against their potential impact in terms of quality, quantity, timeliness and cost:

- Comprehensive terminology resources, proofing tools and translation software tools were deemed to have the biggest potential impact on Quality
- Translation software tools and speech recognition software were deemed to have the biggest potential impact on Quantity
- Translation software tools, speech recognition and comprehensive terminology resources were deemed to have the biggest potential impact on Timeliness and Cost
- Translation software tools and speech recognition were deemed to have the biggest potential impact on Cost

Question 45 & 46 - Open Question on Gaelic and Gaelic SALT

Finally there were two open questions to all respondents to elicit suggestions on what would increase the amount of Gaelic used and any further thoughts relating to SALT. The following is a summary of the 127 responses relating to SALT:[82]

- An interactive online "one stop shop" site for SALT related resources, information and news.
- An online "one stop shop" for all existing and future terminology resources with possibilities for interaction with other users and feedback on terminology issues, including the digitisation of reliable older dictionaries.
- Standardisation of Gaelic technical terminology.
- More use of the International Phonetic Alphabet and audiovisual media in resources to reliably indicate pronunciation to learners and use of technology in a wider context to support the acquisition of good pronunciation
- Better guidance on the use of existing tools, including extremely basic issues, for example, how to disable "Correct as you type" features in Word which turns *tha* into *the* or *i* into *I.*
- Promotion of Research & Development into Gaelic SALT/a Gaelic SALT Research & Development centre that would be able to deal with new challenges and developments more flexibly.
- More attention to be paid to the needs of disabled Gaelic users and their needs for SALT such as screen readers.
- Generally less re-invention, duplication and multiplication of terminology and resources and more co-ordination; for example, the promotion of a single TM software application for Gaelic translators.
- Making existing Gaelic tools (such as spell-checkers and other software) more readily available on public-sector workstations (for council employees, in public libraries, in schools).
- A professional body for translators and interpreters and a qualification in translation/ interpretation.
- A guarded approach to MT.
- Basing more tools online.
- A need for keeping any tool up-to-date.
- A stronger emphasis on open-source tools and development.
- More English to Gaelic terminology resources.
- Gaelic-medium help lines (including technical support for Gaelic software).
- More Gaelic (computer) games.
- A Gaelic thesaurus.

[82] For responses relating to Scottish Gaelic in the wider context, please consult the Attached Files.

Appendix 4

Workshop Reports

All the workshops followed a common pattern and each lasted 3 hours. A professional facilitator[83] ran each event - allowing the researcher to capture feedback and ideas.

The workshops commenced with a short warm-up exercise to get thinking away from traditional discussions.

The second stage was a structured approach using Synectics[84] to answer the Question posed as the workshop topic. Each workshop dealt with a different aspect, selected partially to reflect the interests/expertise of the attendees and overall to give a broad coverage of the problem that was being addressed by the research. These exercises delivered a very rich resource of metaphor for things that could be done and the context where they might be applicable. The topics chosen for this stage were deliberately intended to drive out issues surrounding SALT and its practical usage.

Finally, focus groups were given time to discuss matters they felt important in their professional or everyday use and contact with the Gaelic Language with the intention of raising a short list of things they would like to see implemented or developed. This anchored the thinking into practical applications and outcomes. They were simply posed the question *"Which thing in terms of speech and language technology would make the biggest difference to your work and/or life?"*

Overall the workshops were very successful - not least because they were not talking-shops but were driven to produce workable ideas that could form the basis of strategic and tactical thinking. Some found the approach unusual but interestingly, the attendees generally commented on how enjoyable the experience was, thus adding to their enthusiasm for the output and its potential outcomes.

[83] A practising Certified Management Consultant (CMC) and also a Facilitator with the Open University Business School Postgraduate Programme on Managing Innovation & Change.

[84] A widely-used tool in the field of Innovation and Change to stimulate thinking in a structured and fun way. It uses a "metaphorical process" to make the "familiar strange and the strange familiar". In the experience of the Open University Business School (Europe's largest provider in the delivery of Change and Innovation Management techniques) it rarely fails to deliver across many thousands of instances.

Glasgow

Attendees' background:

Television (MnE)
Adult Education (Stow College), Arts
Actress, Parent of children in GME
Gaelic Officer
Parent of children in GME
Gaelic Tutor in Tertiary Education
Gaelic Translator and Tutor

Creative Workshop Topic

What should a Rules Framework for the Gaelic language look like?

Output

Two main points emerged from the creative session:

Rules Framework

Gaelic requires a rules framework that is independently run. It should:

- Build on existing work carried out in standardising Gaelic.
- Be run by a group of experts in standardisation, (Celtic/Gaelic) linguistics and native speakers of the language.
- Be independent of the education sector.
- Ensure a coherent framework that covers all aspects of the language (grammar, spelling, terminology, etc.)
- Disseminate their work effectively.
- Integrate the wider community into the work as much as possible.

This standard should be promoted for use in formal/technical situations as appropriate.

Preserving the Richness and Diversity

From the outset, the process to develop the Framework should also work towards preserving the richness and diversity of Gaelic - especially in relation to dialectal pronunciation and terminological variation.

At the same time, proper guidance must be given to the users and learners of the language to ensure that people are not left guessing as to which forms are appropriate where and when.

The example used was that of Gaelic terms for "honeysuckle". More than 10 terms for this plant exist. A standard form should be agreed but also include the alternative forms and give guidance on which dialects use these alternatives. It was considered vital that native speakers of the language are not alienated by apparently "forcing" a "sanitised" form of the language on them.

Focus Groups

Media, Government or Private Sector

The first group was composed of professionals using Gaelic on a daily basis in the media, government or the private sector. They would like to see a termbase with the following features:

- Contains new terminology.
- Should be easily accessible, both online and offline (where the suggestion was a downloadable termbase that could be updated via download occasionally.)
- Deals with acronyms, pronunciation of non-translatable terms (e.g. Latin taxonomy, foreign place-names, etc.)
- Gives context and examples of use.
- Aims to re-invigorate and re-introduce native vocabulary that has fallen out of use in preference to inventing new terminology.
- Is located in a single place (the criticism of existing resources was that they are scattered across many locations; it was said that "for place-names and Parliamentary terms you go to the Parliament, for new words to the Stòr-dàta and for usage to Dwelly-d", costing users a lot of time.)
- The termbase should be continuously maintained and expanded by a team of experts and native speakers, ideally (part of) the same group dealing with the Rules Framework.

Gaelic Education Connections

The second group was composed of parents of children in GME and teachers and tutors of Gaelic in further education. It came up with two items.

The first was the concept of a "Blasroom", a resource helping learners with learning good pronunciation and better comprehension of native Gaelic. It should:

- Contain both simple items (like individual words and short phrases) and longer, coherent pieces.
- Deal with a broad range of Gaelic dialects and registers.
- Have a self-checking feature where learners can automatically check their own pronunciation against native pronunciation.
- Be easily accessible and user-friendly.
- Contain a feature where learners outside the Gaelic-speaking areas could experience a live virtual Gaelic-speaking setting, something akin to a real-speech version of SecondLife (see page 162).

The second item was less well-defined but envisaged a tool that would boost the confidence of native speakers in their own speaking abilities and encourage them to use the language much more frequently in various settings. The participants were unsure if speech and language technology would be capable of delivering such a tool.

Edinburgh

Attendees' background:

Translation
University Lecturer
Adult Education
Gaelic Learner, Arts
IT
Media
University Lecturer
Student, IT
Student

Creative Workshop Topic

How the use of Gaelic at the level of tertiary/adult/continuing education be increased?

Output

Aspects of Teaching Gaelic at Tertiary Level

The first point focussed on aspects of teaching Gaelic at this level. Lack of confidence, inhibitions against making mistakes and speaking Gaelic and a mismatch of expectations were felt to be some of the main causes holding back the acquisition of Gaelic. The following measures were suggested:

- Finding innovative ways of putting students at ease when learning/practising Gaelic. Suggestions included whisky, relaxation techniques, nicer learning environments.
- Improving the teaching skills of tutors at this level.
- General measures to combat the widespread overall loss of confidence amongst students towards the end of their first year.
- Increasing awareness at secondary level of the "higher bars" at the tertiary level compared to "fluffy" education at secondary level.
- Making GME less "fluffy" in the final years of secondary.
- Finding ways of encouraging language use outside the classroom.

Mobilising Existing Skills

The second point focussed on mobilising existing language skills at this level of education. It was felt that much of the linguistic talent was not fully mobilised.

- Improving the social networking of Gaelic-speaking students and lecturers across the disciplines through university-based networks such as university intranets and alumni networks. This should include institutions outside Scotland as appropriate, e.g. Nova Scotia, other parts of the UK, Germany.
- Finding ways of encouraging language use outside the classroom.

Developing Higher Registers

The third point focussed on developing higher registers of the language. It was felt that higher registers of the language, including specific terminology, is underdeveloped and in need of development. Given the demographics, it was accepted that this would likely be a graded development as there are currently no degree courses which are not Gaelic-related. Measures suggested were:

- Initially, providing add-on classes for students of other disciplines to improve Gaelic skills relevant to their subject area. Currently, only one university (Aberdeen) is known to hold an informal language workshop for Gaelic-speaking students of law. Such approaches should be formalised, rolled out and encouraged.
- Providing such in cyber-classrooms across Scottish universities to boost numbers.

The MIT Open Courseware project was mentioned, which makes lectures and lecture related material available via download.[85] Such an approach was felt to have enormous potential to deal both with teaching Gaelic at the tertiary level and developing higher registers, perhaps with full access on university intranets and limited taster-access on the internet.

SecondLife[86] University

The idea of a "SecondLife University" was brought up again which could be used to counteract the geographical scatter of Gaelic-speaking students.

General points made

- The culture of not wanting to criticise, not asking hard questions and not pulling up failed projects was criticised as a major obstacle in the development of Gaelic.
- People felt that the wider community was not consulted and integrated into the decision making process enough and that they were rather surprised at the level of consultation in this particular research project.
- Complaints about the lack of communication regarding developments and projects concerning Gaelic.

Focus Groups

People in Tertiary Education

The first group was composed of individuals in tertiary education. Their desired item was a linguistic corpus. This should:

- Be of significant size
- Contain standard tags such as domain
- Should be searchable, including according to domain
- Possibly include spoken corpus material

Also on the list was the production of freely available (not GME-only) teaching resources.

IT and the Media

The second group was composed mainly of people involved in IT and the Media. Similar to the Glasgow group, the desired item was an all-singing all-dancing online dictionary including examples of usage, sound, etc.

A second idea floated by this group was the production of various templates such as letters and forms (e.g. hospital appointment letters, council tax forms, etc.) that could be used both to encourage the use of Gaelic and to avoid wasting resources by separate bodies having to translate the same or highly similar materials over and over.

Both groups felt there was a distinct need for an independent body to regulate and promote the language (a Gaelic Academy), involving linguistic experts, native speakers and the wider community in evaluating and testing, especially of new terminology.

[85] See http://ocw.mit.edu/

[86] SecondLife in this context refers to a particular virtual reality online game (see http://secondlife.com/).

Inverness

Attendees' background:

Translator, former teacher
IT Services
Full-time translator for local council
Comunn na Gàidhlig
Native speaker
Translator

After the icebreaker session, one attendee made his apologies and left, stating he had somehow misunderstood the nature of the workshop.

Creative Workshop Topic

How can technology be used to improve the usage of Gaelic?

Output

Reality for Young People

The importance of Intergenerational Language Transmission is widely accepted but the general feeling is that information regarding "how to" is either not accessible or understandable by the general public. Given the general public is the target audience, the idea was floated that technology could be used to help parents in this respect. More traditional approaches such as an Intergenerational Transmission Hotline (i.e. advice on how to raise bilingual children successfully) were floated but also the radically new idea of producing a Reality TV show that follows a bilingual family through successful language transmission with expert help (along the lines of *Supernanny*).[87]

In passing it was also mentioned that BBC Alba should show more young people not presenting programs but interacting with each other in Gaelic, as this scenario is increasingly rare in real life.

Mobile Technology

Another chief idea was to use GPS and mobile technology to provide Gaelic speakers and interested users with infobytes on their current location. This could range from information regarding the Gaelic history of a place, pronunciation of place-names to local Gaelic-speaking B&Bs or Gaelic events, etc.

Finding a Niche

In terms of technology, it was felt that Gaelic, along with other smaller languages, was always in the process of playing catch-up and/or copying an existing English concept (apart from a few notable exceptions such as sports on BBC Alba or the current affairs programme Eòrpa). It was strongly felt that Gaelic should move away from the catch-up model as much as possible. In order to do so, the following was proposed:

- Invite the Irish and Manx to set up a research and development centre with a Gaelic/Irish/Manx-speaking subsidiary in each country
- Research and develop ideas aimed specifically at the Irish/Gaelic/Manx market. By aiming at a conjoined Goidelic market, the potential market is increased significantly and may even allow for commercial products. For certain products one could even invite other small language groups such as Welsh/Cornish/Breton. To develop

[87] In this context, the recent initiative by Edinburgh University, Bilingualism Matters (www.bilingualism-matters.org.uk) should be a useful potential partner.

commercial models, a best-practice business school should be used but the aim should be a mix of commercial and free products.

- Develop ideas that will not be available in English but that are so attractive even non-speakers will want to use them. A parallel was drawn to the successes of free email services such as Hotmail and Gmail which are highly popular. As a negative example, the inability to switch off the subtitles on BBC Alba was given whereby it was made impossible to consume Gaelic media without the presence of English.
- Whatever products are developed, they will need to be maintained and updated.

Apart from a Celtic email service (which would have to have added value to encourage people to use it), console games (that would be attractive to children) were suggested.

Virtual Learner Environment

The idea of a Virtual Learner Environment was proposed where a limited number of people could experience a fully Gaelic-speaking environment. This could potentially be done in a virtual setting, similar to the idea behind SecondLife but with the explicit goal of producing a totally Gaelic-speaking setting which no longer exists in real life. This would have to be carefully monitored to ensure a working mix of the different speaking abilities. It would also have to be designed so it is extremely easy to use.

This would also be an opportunity to bring Gaelic-speaking employment opportunities to remote areas provided the internet capacities exist.

Focus Group

Due to the small number of participants, the group was not split up for the last exercise.
Points raised in this session were:

- The lack of easily accessible information for translators, such as special Gaelic translators' forums.
- A grammar checker; ideally a sophisticated one but even something that dealt with simple issues for learners/schoolchildren would be helpful.
- Speech to text was deemed useful by Gaelic translators who have to deal with transcribing recordings for various clients. It was felt that even a basic tool might be useful for dealing with large volumes.
- In terms of terminology, it was felt that the higher registers of the language needed clarifying and that there was too much contradiction and variation for technical terminology. Guidance on register usage is also needed. Again it was echoed that it must be made crystal clear to the general public and professional users of the language that such standardised terminology is meant to develop the higher registers of the language, NOT everyday speech.
- At the user end, examples of usage were judged vital.
- It was felt that students, in particular postgraduate students, are an under-used resource in Scottish universities with respect to Gaelic-related study projects. Incentives to encourage students both inside and especially those outside the Celtic/Gaelic departments (in particular IT and linguistics) should be made available (such as study grants) to encourage scientific research into the language.

Skye (SMO)

Attendees' background:

Translation
NHS
Local PA
IT at SMO
Education & Publishing
Translation, Media
Education at SMO
Course Organiser at SMO
IT at SMO

Creative Workshop Topic

How can we create more fluent Gaelic learners?

Output

Focus of Gaelic Development

It was stated and agreed that overall, Gaelic development is focussing too much on education and the media to the exclusion of everyday domains of the language. Graduate placement schemes were criticised for putting students in artificial settings (mostly Gaelic organisations) which reinforce formal domains of the language. As these are the domains students are most likely to be familiar with, the placements do not aid acquisition of informal registers of the language.

It was felt that these schemes should be altered to place graduates with native speakers in settings where they are most likely to encounter informal Gaelic (e.g. with families, crofters, fishermen.) Any such scheme must be to the (economic) benefit of the placement provider.

Perception of Language Learning

The point was also raised that there is a perception that language learning is necessarily an academic endeavour and that this perception needs to be countered.

Standards Implementation

Implementation of standards was deemed to be inadequate. Criticism was voiced that the new spelling rules, for example, were felt to be so prescriptive that native speakers who have grown up using one of the traditional spellings felt they were required to "re-learn" their language. Better guidance is needed on the use of standard language and grammar in formal domains and the acceptance of variants in everyday speech and writing. Frequent changes to any such systems should be avoided and too much emphasis on linguistic purism in colloquial registers should be avoided.

Handling Dialects

Nonetheless, students in adult education need better guidance on how to handle dialects. Some form of linguistic middle ground is desirable in settings where students are unlikely to achieve fluency within a single dialect to avoid frustrating students changing tutors. Tutors should be made aware of the difference between wrong forms and variants and strive towards being tolerant of variants.

Criticism of Isolated Development

Projects such as An Seotal were commented upon as being unhelpful. The development of terminology in isolation was criticised. In the example the lack of additional information was

poor (it was judged "just another wordlist"). The fact it was separate from existing resources (such as the Stòr-dàta) and in particular the fact that the project went live with an extremely small amount of terminological data well below the critical mass for a useful wordlist or dictionary was deemed to be very unhelpful. Moving forward the group agreed that any serious terminology project should also make references to register.

Idiom

In spite of its relatively small size, Faclair nan Gnàthasan-cainnte was singled out for praise for being the only sizeable online tool that lists Gaelic idioms (see Faclair nan Gnàthasan-cainnte).

Among the many challenges facing learners, good idiom was singled out as being particularly weak. At the same time, the personal experience of tutors suggests that most British students are not good at coping with constructive criticism. Apart from measures that improve the acceptance of constructive criticism in general, the idea was floated to use technology to depersonalise criticism. The idea of a tool akin to the *Tip of the Day* feature in various software packages was seen as a potentially effective way of achieving this by offering a Gaelic Tip of the Day add-on that would provide suggestions of idiom, grammar or vocabulary depending on the degree of fluency. A more sophisticated tool might offer context related tips by analysing a text for keywords (e.g. keywords identifying a document being written as a letter or CV).

Pure Gaelic-speaking Centre

It was felt that a purely Gaelic-speaking centre of creativity based in a Gaelic community was desirable to provide a platform for "native ideas" to be created to better reflect the Gaelic view of the world rather than transposing English views and concepts into the Gaelic world via translation. Such a centre should also strive to bring together the older and younger generations to reinforce the Gaelic creativity and to provide access to experience for younger people. The centre should not solely focus on the "Gaelic Arts" but creativity in a broader sense. One target for such a centre should be the development of a "cult concept" to inspire others as there are currently very few Gaelic-speaking cult figures, concepts or celebrities for young people (perhaps with the sole exception of Runrig).

Mobility of Teaching

It was suggested that Gaelic teaching should strive to become more mobile and grow less dependent on traditional "centres of learning". A parallel was drawn to empty churches. Ùlpan was mentioned favourably as it is extremely local as a teaching method. Bearing in mind the seeming reluctance to use new technologies amongst some groups, further possibilities of using such to make learning more mobile should be investigated.

History & Culture

As an aside, it was also commented upon that Gaelic history and culture did not feature prominently enough on the curriculum, both in GME and English-medium education and that, in the absence of changes to the curriculum, outreach programmes are needed to promote a better understanding of Gaelic culture and its role in Scottish history.

Colloquialisms & slang

Apart from the need for developing the formal registers, there is also a need for an online dictionary of highly colloquial language not commonly found in dictionaries (aka Gaelic Street Slang) perhaps with community input.

Random Idea

More interactive Gaelic TV programmes such as a Gaelic aerobics class.

Focus Groups

There were three groupings but overall they coalesced around similar aspirational requirements that included:

- In addition to the Gaelic Tip of the Day (see above), such a tool could also point out common errors.
- A talking dictionary that provides recordings of a neutral pronunciation and some of the more common dialectal pronunciation. Due to the nature of Gaelic morpho-phonology, this should ideally also deal with larger units (e.g. article + noun + adjective) rather than just individual words to demonstrate the sound changes that occur across word boundaries in the language.
- An extended version of Faclair nan Gnàthasan-cainnte with improved search functionalities.
- Predictive texting.
- An updated and improved version of the Stòr-dàta, which remains one of the most frequently and widely-used terminology sites in spite of its limited features.
- A resource site which makes Gaelic teaching tools and resources available to everyone. Many of these tools, in particular those developed within the education sector, are only accessible by mainstream teaching staff. It was felt that making these more widely accessible would benefit the quality of Gaelic teaching overall with little extra effort.
- An online learner's surgery where questions of grammar/idiom/pronunciation would be answered at fixed times by experts in these fields. This needs to involve subject experts to ensure a quality service.
- Digital projects such as Guthan nan Eilean should be expanded and made available more freely on the web to allow maximum access to a wide range of native speaker material speaking about a wide range of topics.

Stornoway

Attendees' background:

IT, Translation
Education, Translation
Education, Translation
Community Education
Media
Educational Publishing
Translation
Translation
Educational Publishing
Translation

Creative Workshop Topic

How can Gaelic translation be improved?

Output

The creative session came to the following conclusions:

Translators' Skills

Although possibly not immediately practical, translators should be encouraged to develop specialisms and move away from being generalists. It is generally the case with bigger languages that translators specialise in certain fields (legal/medical/literature, etc) to ensure familiarity with subject and specialised terminology/phraseology. It was suggested that alongside their general portfolio, Gaelic translators should develop stated specialisms such as "specialising in healthcare".

Mobile Technology

Any technology developed to aid translation should be both future-proofed and accessible via mobile technology. Termbases which are only accessible as PDF file, for example, are virtually impossible to access via mobile technology.

The Workplace

Better guidance needs to be worked out and disseminated both for translators and procurement people. Translators, especially such new to the field, need guidance on how to deal with agencies or procurement people as regards to "the important questions to ask" such as target audience, style, register.

Guidance required

HR/Procurement departments dealing with translation need guidance on the importance of time management as most expect impossible turnarounds (e.g. hurried translations leading tp inferior quality). Most HR/Procurement departments also rarely specify important aspects such as style or spelling and the difference between translation and re-writing impossibly worded original documents.

Need for a Professional body

Overall, it was unanimously agreed that there is a need for a professional, independent body for Gaelic translators and interpreters, a Comann Eadar-theangadairean na Gàidhlig. Such a body should:

- Provide a real/virtual meeting place for translators.
- Serve to disseminate information regarding translation, training opportunities, etc, including a "Translation Ground Rules Fact sheet" for translators and procurement personnel.
- Provide training opportunities for Gaelic translators within the private sector. No specific training courses for Gaelic translation exist. However, it was felt that training in general translation techniques (that are applicable across languages) would hugely benefit Gaelic translation and possibly serve to provide a breeding ground for developing more specialised Gaelic training opportunities in the future.

  Such training could/should take place using web technology given the geographical realities of the Gaelic world. It must also bear in mind the fact that most translators only work in Gaelic translation part-time and may hold a separate day-job. This places certain constraints on the amount of money that translators can spend on technology/training and finding the time to participate in training.

  A training course, if developed following best-practice and aimed at the small language market, could make Scotland a world leader in training translators for small languages.
- Promote the use of idiomatic Gaelic and develop opportunities, real or virtual, that will allow translators and interpreters that have not grown up in a Gaelic-speaking community to connect with "real, idiomatic Gaelic". This could involve specific training courses located in strong Gaelic-speaking areas. Similarly, technology could be developed to suggest appropriate idioms based on English keyword sampling and/or virtual training places.

Terminology

There were urgent calls for Gaelic terminology to be collected and standardised in a single, accessible repository. The current situation where translators are obliged to consult a large number of different sources during translation (often disagreeing on terminology) was deemed highly detrimental to translation, both in terms of quality and workload.

In the development of such a termbase there were also calls for greater attention to examples of usage and the promotion of idiomatic Gaelic, rather than pure (often English-derived) terminology.

Such a termbase should, as a matter of priority, also provide authoritative Gaelic names for standard formulae (e.g. Emergency Exit, No Entry, No Smoking, etc.) and names of bodies and institutions to avoid divergent translations. The lack of accuracy and consistency in some of the recently implemented bi-lingual road-signage in the Isles, Argyllshire and the Highland Region was held up as an example of what can go wrong in the absence of standardisation. This has an obvious negative impact on the Public purse, as well as perpetuating bad Gaelic and negative attitudes towards the language.

Tools

Gaelic speech technology needs to be carefully designed. Overall the experience with English speech-technology has been negative to date, especially in handling Scottish accents and place-names by speech recognition technology. Similar problems with Gaelic regional accents need to be handled carefully.

CAT technology should be designed and disseminated in a way that allows updating thereof and possibly data-collection to expand existing termbases.

Focus Groups

Two groupings produced suggestions as follows:

Translators

There were actually 2 sub-groups but their overall responses are aggregated here.

- An intelligent grammar checker that includes a feature that picks up on repetitive errors and which suggests areas of grammar one may wish to check up on.
- Handwriting recognition.
- An improved version of the Stòr-dàta.
- An independent Gaelic "Academy" of experts and native speakers to oversee the standardisation and development of Gaelic terminology in an open fashion. This should include clarification of advanced grammatical issues for use in formal domains/translations (such as treatment of compound nouns in conjunction with adjectives, use/non-use of the dative case, etc.)
- Better training and guidance on producing idiomatic Gaelic but also simple issues such as keyboard skills and dealing with the accented vowels in various software applications.

IT & Media

- A central site for disseminating information on language tools, accessible to everyone. Even basic information, such as how to handle/enable/disable automatic proofing tools in (Open) Office are not well known.
- A Gaelic thesaurus that is integrated into (Open) Office applications.
- A reverse lemmatiser linked to dictionaries.
- MT software which produces perfectly idiomatic Gaelic.
- A free service should be set up for local/national government bodies and public sector bodies where short Gaelic translations can be proofread by trained experts free of charge. This service should be accessible by various means (email, SMS, web-chat) as long as typographical correctness can be safeguarded.

  Such a service could be located in a Gaelic-speaking area that (currently) has low employment opportunities to raise the economic value of Gaelic in the community.