SALTcymru

Project Closure Report for

SALT Cymru (HE 06 KEP 1002)

Prepared by the
Language Technologies Unit (Canolfan Bedwyr), Bangor University

Rhys Jones, Gruffudd Prys, Delyth Prys, Dewi Jones,
Owain Davies, David Chan, Ambrose Choy

April 2008

PRIFYSGOL BANGOR UNIVERSITY
1884

Llywodraeth Cynulliad Cymru
Welsh Assembly Government

Contents

SALT Cymru Executive Summary

This document describes the activity undertaken by the SALT Cymru (Speech and Language Technologies Wales) project and the key recommendations on the establishment of a SALT Specialist Interest Group in Wales. SALT Cymru has been funded by the KEF Knowledge Exchange Programme, reference HE 06 KEP 1002.

Definition of SALT

For the purposes of the project, SALT (speech and language technology) is defined as the inclusion of human language in software for processing text, speech and knowledge. It includes, but is not limited to the following fields: speech technology; written language input; language analysis, understanding and translation; automatic document processing; machine translation; multimodality; electronic language resources and SALT evaluation.

Economic Importance of SALT sector worldwide

The project establishes that SALT deployments are increasing scales of services, reducing costs and adding to the number of SALT-supported languages. SALT is a significant economic sector: the worldwide market in speech technology deployments alone is estimated to be worth $3.2 billion by 2010. In 2005 the translation market, particularly in multilingual websites and software localization, generated $8.8 billion in worldwide revenues. 4 of the top 20 international companies site their HQs in London, the North of England and Ireland employing over 2300 people.

Potential for development of SALT sector in Wales

The project has discovered a significant appetite for speech and language technologies (SALT) within Wales, both by end users and amongst SMEs that are currently developing them or which have an interest in doing so in future. It concludes that investment, a building up of the knowledge base, and an increase in awareness activities are all required in order that this sector of the economy may capitalize on Wales's privileged position as a bilingual nation and grow to fulfil its full potential in exploiting worldwide markets. It outlines the framework of a specialist interest group to nurture such growth.

The SALT knowledge base in Wales

In terms of SALT development in Wales, the knowledge base is at present relatively small. In Welsh HEIs, it is estimated that the equivalent of fewer than ten full-time academics work directly on SALT. Of these, only about half are permanently contracted to do so. However, there is also a significant amount of SALT associated development work within Welsh academia and industry, where a number of associated SALT developers and companies exploit SALT development together with other technologies. Key developmental points include the growth of multilingualism and multimodality in more complex systems as new ways of interacting with computers.

The SALT Cymru online survey

To investigate which areas of SALT would be of greatest interest to SMEs were they to be developed further, an online survey was set up, suitable for users, developers and potential developers. 48 participants took part in the survey, a higher than anticipated result. Of these 44 stated that they wished to be contacted regarding setting up a specialist interest group in SALT in Wales. The survey results provided a valuable insight into the SALT developed in Wales, with most types of SALT being represented, and no specific SALT type dominating research. Accessibility issues and the linguistic

needs of a bilingual country were highlighted, as were the improvement of interfaces between users and objects or information, and of the identification of what constitutes relevant information to the user, and the meaningful categorization of information.

Further interviews and investigations
Interviews were carried out with individuals in internationally regarded laboratories researching the state of the art in the discipline. These interviews reflect the wide-ranging nature of SALT. Key findings included the importance of multimodality to future developments, and the need to nurture a knowledge base from the earliest opportunity, i.e. in schools in addition to universities. To further investigate the state of the art, the annual LangTech international conference was attended, which provided a broad overview of a wide range of SALT technologies in research and commercial contexts. A focus group with professional translators provided further opportunities to examine the needs of an important group of SALT users and their specific needs and priorities.

Key SALT components
Taking the results of the survey, five key components of SALT were examined. In developing these SALT components, a model is proposed at encouraging development in a non-state aid environment. This involves the use of liberal free software licenses, under which all of the five open software frameworks are currently licensed. The nature of liberal free software licenses means that the then commercially valuable source code need not be released for any end products.

Report recommendations and terms of reference
This report recommends that:
1. A SALT Cymru specialist interest group (SIG) be established. This will coordinate and inform develop the sector in Wales and liaise between academia and industry, and between developers and users of the technology.
2. **The SIG shall have an international watching brief for SALT.** This will bring international best practice to Wales and to promote Welsh interests globally.
3. **The SIG shall work to strengthen the research base in Wales**. It will enable and coordinate funding applications, improve networking and work towards creating a centre of excellence in SALT in Wales.
4. **The SIG shall draw up a prioritized programme to develop a toolkit of basic language resources**. These utilities will be made publically available for download by software developers in Wales.
5. **The SIG shall maintain and further development of the SALT Cymru website to include a resource portal.** It will include provisions for virtual meetings, an on-line newsletter, and details of events of interest to its members.
6. **The SIG shall guide a Welsh SALT education and training programme.** This will provide a clear path from school through HEIs into industry. Use of existing resources will be maximized, with due regard to existing programmes such as KTP partnerships.
7. **The SIG shall address the training and communication needs of its own members.** It will therefore plan a programme of regular seminars and an annual international conference, and use existing events and channels to promote its activities.
8. **The SIG shall address evaluation and quality control issues for SALT developers.** It will draw up recommendations for accreditation procedures and the application of industry standards to the industry in Wales.
9. **The SIG shall seek adequate funding to enable it to fulfil its terms of reference.** It will approach the Welsh Assembly Government to place the SALT SIG on a secure footing and operate efficiently.

1. Introduction

1.1. This report details the activity undertaken between October 2007 and April 2008 by the SALT Cymru (Speech and Language Technologies Wales) project. SALT Cymru has been funded by the Welsh Assembly Government under the KEF Knowledge Exchange Programme.

1.2. SALT Cymru's brief is to advise on the formation, and if applicable to form a specialist interest group (SALT SIG) partnering expert researchers from Welsh higher education institutions (HEI) with small and medium sized enterprises (SMEs) in Wales.

1.3. Speech and Language Technologies (SALT) is a new and emerging discipline. It incorporates expertise derived from human language into an increasing number of computational techniques for the processing of text, speech and knowledge[1]. It is an inherently multidisciplinary area, combining as it does language expertise with information sciences.

2. Context

2.1. In the context of SALT Cymru, Wales is seen as a living bilingual and multilingual laboratory for the identification and exploitation of new opportunities. An opportunity exists, therefore, to bring Wales' existing research expertise in SALT together with key players, interested parties and potential markets. Through these, coherent initiatives will be developed.

2.2. SALT Cymru is a particularly timely project. The United Nations has designated 2008 as the International Year of Languages, and Bangor University will host a major conference in August 2008 as part of its celebrations[2]. SALT Cymru therefore comes at one of the most ideal times for Wales to show what it has to offer in this area.

2.3. SALT-related activity within the European Union has previously received funding within the scope of knowledge economy initiatives such as Interreg and the Framework Programmes. Wide European collaboration is inherent in such funding schemes, leaving national governments with the responsibility to fund internal research and development in their own countries. In the context of European participation, SALT Cymru may be considered as useful preparatory work which may lead to larger, transnational projects in the longer term.

3. Objectives

3.1. The stated aims and objectives of SALT Cymru are as follows:

- To investigate the current SALT research base in Wales
- To identify potential key players in academia, public organisations and industry

[1] The component parts of the field are also sometimes grouped as *human language technology*, *language engineering*, or *applied computational linguistics*. In this document, the acronym SALT, for *speech and language technologies,* will be used to describe these.

[2] The conference is titled Global Understantding in Multilingual, multimodal and multimedia contexts (GUM3C), see website http://www.gum3c.org/

- To analyse present state of play and needs of industry in Wales, with special reference to SMEs and their markets/potential markets
- To describe international best practice in SALT, including cutting-edge developments and forthcoming innovations in multilingual multimedia applications
- To evaluate relevant open-source software and standards for use in Wales in a pre-competitive research environment
- To pilot use of focus groups, networks, seminars, web site and other consultation and dissemination methods for specialist interest group communication
- To draw up recommendations and terms of reference for the future work of the SALT Cymru specialist interest group

The SALT research base in Wales

4. Preamble

4.1. SALT Cymru is primarily a technology transfer project, with the aim of setting up a specialised interest group (SIG) to bridge the gap between Wales' SALT research base and users/potential users. Awareness of the scope and scale of the SALT research base in Wales is a prerequisite for any such activity to take place.

4.2. To define the SALT research base in Wales, we must first define what SALT itself is. This presents challenges. SALT is considered in academia to be a multidisciplinary area, hence it is not neatly constrained by the boundary of one subject or even multiple subjects. Depending on the definition of SALT taken, the discipline can involve expertise in computer science, electrical engineering, linguistics, media and communication studies, psychology and creative arts. Even this list is not exhaustive.

4.3. An important first step in the SALT Cymru project was, therefore, to present an adequate definition of SALT. This would enable easier classification of users, developers and potential developers.

5. Defining speech and language technology for SALT Cymru

5.1. Appendix A presents a representative sample of mission statements and definitions of work programmes, gathered from academic departments undertaking SALT work in the UK and further afield. At first sight, then, it might appear that a workable definition of SALT might be achieved by merging this sample into one. This approach is not without its problems, however. Each university department will define SALT to suit its own ends. It is not possible to guarantee that an individual research unit or department will produce an inclusive definition of SALT; neither is it possible to guarantee that the combined definition will similarly be inclusive.

5.2. Rather than attempt to define SALT from others' definitions, use is made of an authoritative overview of the field, published by a group of researchers of international standing as *Survey of the State of the Art in Human Language Technology* (Cambridge University Press). First introduced in 1996 by and currently being revised, it presents developments in each field of SALT, therefore giving a de facto definition of what SALT entails.

5.3. The *Survey of the State of the Art* divides human language technology (i.e. SALT) into the following thirteen categories:

1. Spoken Language Input (speech recognition, speaker recognition)
2. Written Language Input (OCR, handwriting recognition)
3. Language analysis and Understanding (grammar formalisms, semantics, parsing)
4. Language Generation (syntactic generation, deep generation)
5. Spoken Output Technologies (text-to-speech techniques)
6. Discourse and Dialogue (spoken language dialogue modelling)
7. Document Processing (text extraction, interpretation, summarization)
8. Multilinguality (machine translation, translation aids, multilingual information retrieval, multilingual speech processing)

9. Multimodality (gesture and facial movement recognition, visualisation)
10. Transmission and storage (speech coding and enhancement)
11. Mathematical methods
12. Language resources (written and spoken corpora, lexica, terminology)
13. Evaluation (of all of the above)

5.4. Whilst comprehensive, it is felt that for ease of a definition, the thirteen categories can be reduced in number. Speech input (category 1) and output (category 5) can be combined under the banner of speech technology. The more linguistically aligned sub-disciplines of language analysis, understanding and generation (categories 3 and 4) can similarly be combined with discourse and dialogue (category 6). It can be further argued that 'mathematical methods' (category 11 above) and multilinguality (category 8) are inherent in many of the fields that make up SALT. A process such as this reduces the 13 categories to 8, as follows:

1. Speech technology (speech recognition, speaker recognition, text-to-speech techniques, speech coding and enhancement, multilingual speech processing)
2. Written language input (optical character recognition, handwriting recognition)
3. Language analysis, understanding and generation (grammar, semantics, parsing, discourse and dialogue)
4. Document processing (text and term extraction, interpretation, summarization)
5. Machine translation (including computer-aided translation, multilingual information retrieval)
6. Multimodality (gesture and facial movement recognition, visualisation of text data)
7. Language resources (written and spoken corpora, lexica, terminology)
8. Evaluation (of all of the above)

5.5. Therefore, the final definition of SALT derived from the *Survey of the State of the Art* is as follows:

> SALT (speech and language technology) is defined as the inclusion of human language in software for processing text, speech and knowledge. It includes, but is not limited to the following fields: speech technology; written language input; language analysis, understanding and translation; automatic document processing; machine translation; multimodality; electronic language resources and SALT evaluation.

This is used as the benchmark definition for SALT Cymru.

6. Mapping the SALT field: developers, associated developers and users

6.1. Having produced a definition of SALT, a distinction can now be drawn between the following categories of people:

- SALT developers
- SALT associated developers
- SALT users

SALT developers

6.2. This term is fairly intuitive in its definition: it encompasses all those who develop SALT as part of their work. The term 'developer' here is used in preference to 'researcher', to include not only those in academia, but also those in industry developing speech and language technology.

SALT associated developer

6.3. There are many fields that are not directly related to SALT, but which make use of speech and language technology in applied research work. There are other research areas, particularly in signal processing, language and linguistics, whose results form a basis for developing SALT itself.

6.4. We use the term *SALT associated developer* to describe developers in both these situations. As will be seen in later sections, pure SALT development work in Wales seems to be relatively fragmentary, geographically dispersed, and in general is accomplished by individuals or small teams. Associated SALT development within Wales appears to have a somewhat stronger research base.

SALT users

6.5. At first sight, defining users of speech and language technology might appear to be a straightforward process. There is, indeed, little that is complicated in defining users of speech and language technology given a clear definition of what that technology is. However, a problem arises when we consider whether SALT users would define themselves as such.

6.6. For instance, the definition of SALT in Section 5.5 is broad enough to include spelling and grammar checkers as speech and language technology, and similarly with scanning software which uses optical character recognition. Yet few of those actually using, say, a spell checker would think of themselves as SALT users.

6.7. It is apparent, therefore, that the process of determining who constitutes a SALT user has to be an objective one. It is often not productive to arrive at such a determination simply by asking individuals whether they perceive themselves as 'SALT users', given that the technology is transparent. The question of how users should be made aware of the technology they are using, or indeed whether this should happen at all, is addressed later in this document.

7. The SALT research base in Wales

7.1. With SALT defined, and the divisions of organisations and individuals into 'developers', 'associated developers' and 'users' having been established, a survey can now be accomplished of the SALT research base in Wales.

7.2. To accomplish this survey, a list was drawn up of university departments that might conduct SALT development or associated SALT development (see Appendix B). A search was then made via the websites of all higher education institutions in Wales, via other web-based resources detailing academic research, and also via printed materials such as research magazines and periodicals.

8. Areas of SALT Research and Development in Wales

8.1. As a multidisciplinary field, it is perhaps inevitable that SALT research should suffer from fragmentation across several academic departments, leading to isolation of researchers. This is particularly true in Wales, where SALT and undertaken by computer scientists, electrical engineers, applied linguists and terminologists. Whilst these are largely aware of each others' presence, little or no concerted collaboration has taken place between these departments.

8.2. The investigations of SALT Cymru concluded that the SALT research base in Wales is small and fragile.

9. SALT research in Welsh academia

9.1. Appendix B presents areas of SALT research in Wales' higher education institutions.

9.2. Following a survey of all relevant departments in higher education institutions in Wales, only three academic institutions in Wales were found to undertake pure SALT development.

9.3. The institutions undertaking SALT development were:

- Swansea University, where one full-time academic, supervising six PhD students, conducts work on speaker recognition, biometrics and related fields
- University of Wales, Lampeter, where a small team of two or three full-time researchers previously maintained an on-line Welsh-English dictionary
- Bangor University, where
    - a team of eight researchers (five full-time and three part-time) develop technology including speech synthesis, spelling/grammar checkers and electronic dictionary resources
    - a team of two researchers undertake work on text and document summarization and intelligent data mining as part of the Knowledge Discovery Research Group
    - a team of two or three researchers are engaged in language visualization as part of the Visualization and Modelling Research Group.

9.4. It is estimated, therefore, that the equivalent of fewer than ten full-time academics work within Wales on SALT. Of these, only about half are permanently contracted to do so.

9.5. It is important to note that there is also a significant amount of SALT associated development work undertaken in Welsh academia, the most notable of which includes:

- The Digital Signal Processing Centre, part of Cardiff University. Digital signal processing (DSP) is not considered part of SALT as such, but widespread use is made of DSP techniques in some of the low-level processing required for speech technology
- A researcher in Swansea Metropolitan University similarly has interests in DSP techniques
- Linguistic research in Swansea and Bangor universities. Swansea University contains the Centre for Applied Linguistic Studies (CALS) which has in the past

developed computer software to assess the vocabulary of learners of English as a second or foreign language. Bangor University contains the ESRC Research Centre for Bilingualism in Theory and Practice.
- Research on multimodality in the Future Interaction Technologies Lab at Swansea University. The work accomplished in the Swansea lab is not part of SALT as such, but developments in multimodal technologies are increasingly informing and driving developments in SALT itself.
- Research in Glamorgan University on the use made of minority language computer software, and related speech and language technologies, by the speakers of those languages.

10. SALT research in Welsh industry

10.1. In industrial settings, there are few companies that specifically develop SALT in Wales. By the definition of SALT Cymru, and to the best of the project's knowledge, there are no commercial organisations in Wales engaged in pure SALT research and development. There are, however, a small number of associated SALT developers in Wales. Through a combination of directory searches and the personal knowledge of the project team, the following associated SALT developers were found to operate in Wales:

- Geolang, a two-person business in Pembrokeshire, which specifies ISO language codes for the world's languages, and undertakes other standardisation work in linguistic and terminological fields
- Ensigma, a company in Chepstow, which designs digital signal processing hardware for digital radios (sold to the consumer under the brand name Pure). It is part of the larger Imagination Group.
- Megabee, a four-person business in Monmouth, which develops a hand-held writing tablet to aid frequent communication (by carers) with patients who have no, or impaired, ability to speak and who cannot write legibly. Speech output is in the process of being included in the Megabee device, the module being bought in from a third party outside the UK.

10.2. This relatively small number of organisations and fragile research base is surprising given the global importance of speech and language technology. In 2005, the automotive speech technology sector alone, comprising one specific application of a sub-section of SALT, was estimated to be a $4.4 billion industry by digital technology consultants Strategy Analytics. SALT is a significant economic sector: the worldwide market in speech technology deployments alone is estimated to be worth $3.2 billion by 2010. In 2005 the translation market, particularly in multilingual websites and software localization, generated $8.8 billion in worldwide revenues. 4 of the top 20 international companies site their HQs in London, the North of England and Ireland employing over 2300 people.

11. Analysis of the SALT research base in Wales

11.1. In terms of a SALT research base, the activities undertaken in Wales appear to show considerable breadth of development. They encompass both speech and language technology, and of the eight research areas defined by the project as part of SALT, seven are researched within Wales, namely:

- Speech technology (speaker recognition in Swansea, text-to-speech techniques and a speech recognition project in Bangor)
- Language analysis, understanding and generation (spelling and grammar checking, and a library of language tools in Bangor)
- Document processing (text and term extraction, in Bangor)
- Machine translation (a small pilot project in Bangor)
- Multimodality (SALT-associated research in Swansea)
- Language resources (developed in Bangor, Swansea and Lampeter)
- Evaluation (standardization work in Bangor and by Geolang, undertaken with ISO)

11.2. It is also noted that there is very little duplication of subject areas within Welsh SALT research. It is rare to find multiple companies or academic institutions active in the same sub-field of SALT, as each group seems to have its own individual niche. This finding is likely to be a positive one for any specialist interest group; a lack of conflicts of interest within the group is likely to lead to a more receptive environment for sharing ideas, and working together in a collaborative environment. The model of open innovation (see Appendix J) is also of interest in promoting cooperation in the SALT sector in Wales.

12. Interviews with key players

12.1. To gain a greater understanding of the SALT research base and associated research base in Wales, interviews were carried out with individuals that it was felt might be able to illuminate current developers in their fields and provide pointers to future developments. A broad cross-section of individuals was chosen, and care was taken to promote a wide geographical base of respondents (in north and south Wales) in addition to a wide range of interests.

12.2. To this end, the following individuals were chosen for interview:

- Dr Matt Jones, Reader at the Future Interaction Technologies Laboratory, University of Wales Swansea, chosen for his knowledge of how SALT technologies might integrate into the wider arena of computer-human interaction
- Dr Daniel Cunliffe, Senior Lecturer in Multimedia Computing at the University of Glamorgan, chosen for his expertise in how SALT is used in minority language cultures, particularly in the Welsh context
- Tegau Andrews, a researcher in the Modern Languages Department of Bangor University, chosen for her knowledge of translation technologies in Wales, how they are deployed and used in industry, and how language-driven industries in Wales might benefit from such technologies where they are not currently being used
- Richard Sheppard, Managing Director of Draig Technology Ltd, chosen for his experience of using SALT and other computer-based technologies in industry, and for his awareness of the needs of Welsh SMEs in these areas
- Delyth Prys, leader of the Language Technologies Unit at Bangor University, chosen for the range of SALT related activities undertaken in the unit.

Detailed notes from all the interviews may be found in separate sections of Appendix C.

12.3. The interview with Dr Matt Jones (Appendix C1) is particularly useful in placing SALT technologies in their wider academic and industrial context. SALT is seen by his laboratory as a means to an end, rather than an end in themselves. The laboratory of

which he is part deals in multimodal interfaces to computers. Multimodality, in computer science, can comprise anything which moves human interaction with computers beyond the now traditional keyboard/mouse interfaces. Therefore, speech input and output would be seen as one possible mode of human-computer interaction amongst many others. This points to the need to view SALT not only as a worthwhile development in itself, but as a building block to other technologies. These technologies may be regarded as 'blue-sky' developments at present, but it is likely that some of them will find mainstream acceptance in due course.

12.4. The interview with Dr Daniel Cunliffe (Appendix C2) reveals a gulf between the development of SALT and its uptake by users. This evidences itself in two forms. Firstly, users are not aware of the range of SALT applications that have been developed and that are available to them. Additionally, even those users that are aware of those applications may place too much faith in their infallibility. The specific example given of the latter case is of a Welsh translation of an interface to a search engine: this may be used by web searchers without their being aware that it does not contain all the functionality that might be expected from a full Welsh localization, such as the ability to recognise and process mutated and plural forms of words.

12.5. The interview with Tegau Andrews (Appendix C3) highlights the lack of uptake of SALT tools in one particular section of Welsh industry, namely computer-aided translation (CAT) tools amongst Welsh translators. Such tools are seen as essential aids in the international translation industry, aiding productivity, quality control and management of documents and projects. In analysing the reasons for the low uptake of such tools, a number of factors were found. Lack of awareness of the increased efficiency achievable through the tools was one of these factors; but another was the lack of training available in the use of the tools themselves. This points to education and training in SALT as essential in improving not only the uptake of SALT in Wales, but also the confidence of users in realising the potential of SALT in their day-to-day work.

12.6. The interview with Richard Sheppard (Appendix C4) discusses the growth potential for SME developers of SALT in Wales. It is stated that this can happen in two main ways: firstly, by their developing solutions for a Welsh bilingual market; secondly, by developing multilingual capacities for broader, worldwide markets. Richard Sheppard does however feel that the potential for SALT in Wales currently remains as potential. He points to the need for having marketable case studies for SALT to illuminate business decisions. He mentions deficiencies in training in computer technologies generally, whether that training is provided by academia or by the private sector. Further, Richard Sheppard states that there needs to be greater understanding both in the public and private sectors of what SALT can offer them, leading to embracing by businesses and organisations of the potential of the technologies. For nurturing future development in SALT, he suggests that it would be valuable to create a basic suite of language tools that SMEs can customize and integrate into their own products. His idea for a web-based portal for language resources is an innovative idea that merits consideration by any new special interest group and is further discussed in the recommendations.

12.7. The interview with Delyth Prys (Appendix C5) gives an account of the activities of the Language Technologies Unit at Bangor University. This was undertaken as an additional internal exercise in order to demonstrate the range of activities undertaken by the unit. Of the eight SALT categories identified in Section 5.4 of this report, only one of them is so far not covered by their activities. Further, the unit provided a workable model

of joint co-operation between university departments in a project which delivered basic speech resources for both Welsh and Irish. It is currently active in further developing applications which have been licensed to many Welsh SMEs on a B2B (business to business) model, and in a Knowledge Transfer Partnership project to develop Welsh speech recognition and machine translation both within the associate company and for potential sale to other users.

The SALT use and needs of Welsh SMEs and individuals

13. The SALT Cymru survey

Methodology

13.1. A key part of SALT Cymru was to investigate the use of speech and language technology by Welsh SMEs, and to see which areas of SALT would be of greatest interests to SMEs and other users were they to be developed further. This investigation is primarily accomplished via a web questionnaire. A focus group comprising consultations with others interested in SALT was also carried out, and is described in Section 16.

13.2. In developing the web questionnaire, a search was made for suitable packages that would accommodate multilingualism. Limesurvey[3] was found to provide a solution. It was available through a free software license, and for the purposes of the project the package was translated into Welsh. The resulting Welsh translation is being made available to others through the Limesurvey website. The result of the translation was that while filling in the questionnaire, users could opt to view the questions and answer either in Welsh or English, and the results could then be aggregated by the SALT Cymru team.

13.3. To gather participants for the web questionnaire, three main methods are used. They are as follows:

- The distribution of invitations to participate to via press releases, which are publicised in media including:
    - The North Wales Daily Post
    - Advances Wales, a Welsh Assembly Government publication devoted to Welsh innovation, read by a significant number of high-tech companies
    - The news section of IT Wales, aimed at IT professionals both in industry and academia
    - The UK Research Council website, thus reaching academic-linked companies
- Promotional literature such as leaflets and posters prominently displaying the website address, and marketing of both the website and survey at exhibitions and conferences attended by the SALT Cymru team.
- The sending of personal invitations to known participants in SALT research and development. This includes those mentioned in previous sections of this report, and those who have previously approached Bangor University's Language Technologies Unit for consultations regarding SALT development and deployment.

13.4. The survey was not made available to the general public, and hence a unique token was given to each participant. Because of this, most respondents go through a two-stage process: first, of registering their interest via the project website; then, of receiving an email containing a link which contains their unique token, allowing them to complete the survey. Those receiving a personal invitation due to their known interest or involvement in SALT do not go through the initial registration process: tokens are generated for them in advance and they receive a direct link to the survey by email.

[3] http://www.limesurvey.org/

13.5. The web questionnaire is designed to be straightforward to complete, thus maximising completion rates. It avoids excessive length and is kept as straightforward as possible. Questions rendered irrelevant to the respondent by their earlier answers are not displayed, and guides and definitions accompanied questions to clarify and explain the terminology used. Survey respondents may save their responses mid-session and return to them at a later date, and the respondents can participate anonymously if desired.

13.6. The web survey divides participants into three categories, as previously defined in this report: SALT users, developers and potential/prospective developers. At the start of the survey, participants are asked questions which will determine into which category or categories they will be placed. This reduces survey completion time, as subsets of questions are then asked in order to target each category of users in turn.

13.7. In developing the survey, it was decided to minimise as far as possible the use of free response questions, i.e. those where respondents could enter unrestricted text in answer to a question. This was done in order to reduce survey completion time for the respondents, and also to enable trends to be detected as clearly as possible: this can be done more simply in multiple choice questions than in free response ones, as qualitative analysis is not necessary for the former.

13.8. A full list of survey questions can be found at the start of Appendix D.

14. Survey results

14.1. A total of 48 respondents completed the survey. This is considered to be a reasonable number, taking into account the specialized nature of the survey and the project's scope and timeline.

14.2. As expected, the respondents to the survey all declared an interest in SALT. Of the survey participants:

- 40% stated that they were SALT developers
- 31% stated that they were prospective SALT developers, who might be interested in developing such technologies in the future
- 29% stated that they were users or prospective users of SALT: they did not develop SALT and had no intention to do so in the future, but were interested in the technology from a user perspective

The proportion of users stating that they have an interest in developing SALT in the future is encouraging, as this points to the potential growth of the discipline.

14.3. The full survey responses, including a breakdown of responses to each question, are given in Appendix D2.

14.4. The detailed responses will not be repeated here. It is felt that the key questions in the survey are those which aim to discover the specific SALT interests of the respondents, be they users, developers or potential developers. The findings in these areas are described in the following sections.

Developers' interests in SALT

14.5. SALT developers were asked which SALT products they developed, or had developed in the past. The most popular areas that had been developed were as follows:

- 19% developed speech enabled communication aids for disabled users and those with specific needs
- 15% developed speech recognition by computer (dictation software)
- 15% developed text proofing tools (spelling, grammar and language checkers)
- 15% developed intelligent web searching techniques
- 9% developed software involving keyword spotting and trend spotting from text
- 9% developed machine translation software

14.6. SALT developers were also asked which aspects of SALT were of greatest interest to them. This information was gathered by asking developers to rate how important they considered various aspects of SALT. Of these, the following aspects were considered 'very important' or 'fairly important' by the greatest numbers of respondents:

- 78% for text proofing tools
- 74% for speech-enabled communication aids
- 68% for intelligent web searching
- 57% for automatic translation
- 52% for speech recognition by computer
- 47% for keyword/trend spotting from text
- 37% for text-to-speech systems

Potential developers' interests in SALT

14.7. Potential SALT developers were asked which areas of SALT they might develop in future. This question was asked in order to be able to map and possibly prioritise future activity by any SALT special interest group that might be formed. The five most widely quoted areas for future development were as follows.

- 16% might develop methods to aid written language input
- 12% might develop language analysis tools
- 12% might develop machine translation
- 12% might develop speech technology
- 9% might develop multimodal techniques

Users' interests in SALT

14.8. In order to determine whether developers' and potential developers' interests in SALT mirrored those areas of SALT of interest to end-users, those who described themselves as SALT users were also asked what areas of SALT they used most regularly. In the results below, 'regular use' is defined as using the specified SALT component once a month or more frequently.

- 96% regularly used text proofing tools (spelling/grammar checkers)
- 85% regularly used electronic language resources (such as online dictionaries)
- 35% regularly used OCR (optical character recognition) software
- 36% regularly used computer-aided translation (translation memory) software
- 32% regularly used other language analysis software
- 29% regularly used machine translation software

- 25% regularly used text-to-speech software
- 22% regularly used speech recognition software

This represents a wide spectrum of SALT use, which is encouraging for nurturing broad-based development within the sector in Wales. It is not surprising to see such a high percentage of use of spelling/grammar checkers and online dictionaries, given their low cost (usually free with word processing software, or free for use on the web) and ease of use by non-specialist users. The comparatively high usage figure for computer-aided translation can be explained by publicity given to the SALT Cymru project at the Association of Welsh Translators' conference. This meant that more of the respondents were involved in translation activities than might be the case for SALT users in the general population.

Other developments by SALT developers

14.9. In developing the survey, it was appreciated that companies might not concentrate solely on SALT development, especially as the market for SALT in Wales has not yet been fully developed. Consequently, it was decided to ask developers in which other areas of research, if any, they participated. 79% of respondents developed some technology other than SALT. The results for all respondents who answered this question are shown below: as multiple responses were possible for this question, the percentages total over 100%.

- 58% developed computer software (excluding web technologies)
- 47% developed web technologies
- 42% developed telecommunications technologies
- less than 11% each developed biotechnologies, visualisation technologies, biometric technologies, or industrial sensors/systems

Funding streams

14.10. In order to determine possible funding streams for future SALT research, those who were interested in developing SALT were asked about their main source of funding. Of those who could specify their main funding source:

- 37% were mainly funded by private finance
- 21% were mainly funded by central government (whether from UK central government or the Welsh Assembly Government)
- 14% were mainly funded by research grants (from UK or European research council)

The prevalence of multiple funding sources for potential SALT developers is encouraging, as it means that SALT development in Wales is not entirely dependent on one main stream of finance, whether that is private or public. It should be noted, however, that the amount of funding per company is not queried: it was decided beforehand that most organizations answering the survey would probably not be willing to divulge this information. It is possible, for example, that one large research grant awarded to an academic institution might dwarf several smaller governmental grants given to a larger number of smaller companies or organizations. Hence the information above represents the numbers of developers receiving funds from various streams, and does not necessarily reflect the relative amounts of finance received via each stream.

Further, due to its dependency on research grants as its major funding source, higher education research is vulnerable to fluctuations in its grant-capture success rate. The foundation of permanent centres of SALT research in Wales should be seen as a priority if the future of SALT research in Wales is to be safeguarded.

15. Key findings from survey

15.1. The SALT Cymru survey detected a pleasing interest in SALT in Wales. The survey respondents, 48 in number, were higher than expected. All showed some degree of interest in SALT, whether as users, developers or potential developers: this is to be expected given the degree of self-selection amongst the survey's respondents.

15.2. While a wide variety of SALT areas were mentioned by respondents, some appeared to be of consistent interest to developers, potential developers and users. These included:

- Text-proofing tools
- Speech-enabled technologies (whether speech synthesis as a module, or speech-enabled communication aids as finished products)
- Speech recognition
- Intelligent web searching
- Keyword and trend spotting from text
- Machine translation

The survey results indicate that the SALT developed in Wales is varied, with most types of SALT being represented, and no specific SALT type dominating research. This is to be welcomed as it demonstrates that SALT development in Wales is not overly dependent on any one SALT category.

15.3. Extrapolating from Section 15.2, it would appear that SALT development in Wales caters for:

- the accessibility needs of disabled users
- the linguistic needs of a bilingual country
- the improvement of interfaces between users and objects or information
- the improvement of the identification of what constitutes relevant information to the user
- The meaningful categorization of information

Because of SALT's suitability to address many of the needs of disabled users, a large proportion of work in the field of SALT is to do with improving accessibility. The existence of government legislation to ensure accessibility provision for disabled users provides an added economic stimulus for this sector. Work to adapt SALT-based accessibility aids for disabled users who wish to use Welsh is known to be in progress, and is an area with great potential for co-operation between sectors such as Higher Education and Not-for-profit Organizations.

15.4. The awareness of SALT and perceptions of SALT varied considerably amongst survey respondents. This might not have expected to have been the case; the survey group, comprised as they were of people who had opted to complete a questionnaire in

this technical field, would reasonably have been expected to be self-selecting and aware of the areas in which they would be answering questions. However, despite comprehensive and helpful definitions (as seen in Section 5.5) alongside survey questions and on the SALT Cymru website, there was still considerable confusion over what constituted SALT. In particular, some respondents to the survey did not consider themselves to be SALT users, yet it can be reasonably expected that they would use spelling and grammar checkers (considered part of SALT according to the definition) on a regular basis. While SALT Cymru is not directly concerned with educating users in what SALT is, the issue of how to communicate to users what their SALT needs might be should be borne in mind. It is discussed in more detail in Section 23, which considers the role of the special interest group.

15.5. It appears, both from the results of the survey and the findings of Sections 9-11 of this document, that Welsh academia provides a greater range of SALT development than businesses. This is to be expected, as SMEs in particular must usually concentrate on a specific aspect of technology in order to develop a finished product, within their constraints of manpower and funding. The more fundamental building blocks for SALT are developed by academia, and SMEs can in some cases then make use of this in their product. The issue of licensing academic developments in SALT for the benefit of SMEs will be discussed further in Section 20.8.

15.6. Experience garnered in the development of language resources for Welsh, and the creation of the corpora and lexica that lie at their heart has resulted in a base of SALT expertise within Higher Education which is could be further exploited, especially through strengthening connections between HEIs and industry through models such as Knowledge Transfer Partnerships.

15.7. For similar reasons, it is noted that many SME SALT developers do not solely develop those technologies. 79% of SALT developers surveyed stated that they also developed other technologies within their organisation. The most common technologies to be developed were computer and web technologies, reflecting the high degree of computing knowledge required to develop many SALT techniques.

16. Focus group

Methodology

16.1. It is believed that the survey presents a valuable overview of the needs of Welsh users and developers in SALT. However, it was also felt that significant groups of businesses and users might be unaware of what SALT might have to offer them, and so might not think of completing the survey. It was therefore decided to convene a focus group of users in an industry which had a high potential of benefiting from SALT developments.

16.2. The Welsh translation industry is one whose members process large amounts of speech and language on a daily basis. It is also numerically strong. There are over 220 addresses on a Welsh terminology mailing list, who are mainly professional translators[4]. The Association of Welsh Translators and Interpreters[5], a body where membership is dependent on successfully passing examinations demonstrating professional competence, has about 140 members. These include translators working in both the public sector, e.g. for local authorities in Wales, and also private translation companies. Significantly these companies are often SMEs based in rural areas of Wales, providing good quality employment at graduate level, and employing a high percentage of women.

16.3. Due to the numerical strength of the Welsh translation industry and its particular potential for using SALT to increase its efficiency and productivity, a focus group was convened as part of the Association of Welsh Translators' annual conference in Aberystwyth on 16 November 2007. It was part of the main stream of the conference, and about 80 people were in attendance.

Findings

16.4. The result of the focus group discussions are given in detail in Appendix E1. The main findings are summarized below.

16.5. Initially, it was found that there was little knowledge of SALT within the focus group. Fewer than 50% of those present used translation memory software, a finding echoing the more detailed survey of translators carried out by Tegau Andrews (Section 12.5). However, following practical and clear explanation of various components of SALT, the level of interest shown in various aspects of the technology was seen to increase within the group. This points to a need to explain such technologies in informal, easily accessible ways, a finding which has implications for the work of any special interest group to be set up in the field.

16.6. Perhaps predictably, greater interest was shown in the available technologies if they could be made available at low or no cost. This may not be a viable option for SMEs producing SALT. It may however be reasonable for academic institutions developing SALT under centrally funded research grants to release the results publically without cost (in many cases, this is a condition of grant award). Further discussion of

[4] Welsh-termau-cymraeg (run by the Language Technologies Unit at Bangor University), see http://www.jiscmail.ac.uk/lists/welsh-termau-cymraeg.html
[5] For more information, see website http://www.welshtranslators.org.uk/

how free products and modules developed by academia may be viably exploited and monetized by industrial organizations may be found in Section 20.8.

16.7. A number of technologies mentioned by the focus group were relatively simple ones. For example, in setting charges for their work (normally charged per multiple of words in the source texts), translators would benefit from word counting software which could import Microsoft Word (.doc) and Adobe Reader (.pdf) files, and which could also count material in text boxes. While software such as this is peripherally related to SALT, it would not normally be considered as part of SALT itself, and it is significantly simpler than much of that which is developed as SALT. It could be argued, however, that taking development time into consideration, such a module might well be of greater benefit to improving translators' efficiency than more complex software.

16.8. In terms of specific areas of SALT, the focus group had specific interests in speech recognition, translation memory software and optical character recognition (OCR) software. This reflects the common workflow of a translator, where documents must be initially processed (requiring scanning in some cases), translations must be consistent (necessitating translation memory software) and a significant amount of typing is required in producing the translations (a process which would be aided by speech recognition for the target language). The needs of disabled workers were also discussed, especially in terms of speech recognition software.

Conclusion

16.9. The focus group has presented a number of pertinent conclusions for any specialist interest group activity. It has underlined the importance of clear exposition of SALT to any potential users of the technologies. It has also underlined the potential of the speech and language technology market in Wales, provided care is taken not only to provide tools that match the requirements of a user group, but also to explain the potential benefits of the tools to the users themselves.

17. The state of the art in SALT

17.1. An important part of the SALT Cymru programme is to map current international developments in speech and language technology, in order to ensure that any developments in Wales are compliant with the state of the art. Further, it is hoped that it will be possible to highlight current research which is suitable for exploitation in the Welsh context, either through modules to be developed by academia, or as products using these modules that can be sold by businesses as finished products.

17.2. In order to pinpoint those areas of SALT that are of most interest to potential users, the results of the survey and focus group are taken. Section 15.2 and 16.8 show that the key areas of interest are as follows (presented in no particular order):

1. Text-proofing tools
2. Speech-enabled technologies (whether speech synthesis as a module, or speech-enabled communication aids as finished products)
3. Speech recognition
4. Translation memory systems
5. Machine translation
6. Intelligent web searching
7. Keyword and trend spotting from text
8. Optical character recognition

17.3. In the case of text proofing tools (spelling and grammar checkers), it is felt that modules for these are already sufficiently mature in the English and Welsh languages. Spelling checkers exist for both languages in Microsoft Office[6] and in OpenOffice[7]. In the case of English, a grammar checker is included as part of Microsoft Office and a free grammar checker is also available which integrates with OpenOffice[8]. A comprehensive grammar checker for Welsh is available as a standalone product[9], and a free online grammar checker is also available[10]. However, the later is deemed to be of low quality, being the work of an enthusiastic amateur. Open-source software can be variable in quality: while some is quality-controlled and highly reputable, the voluntary nature of the activity means that other software may not attain the same standard and may not be suitable for professional use.

17.4. Text proofing tools are largely seen as ‘free’ utilities. This is of course illusory: the cost of developing and integrating spelling/grammar checkers for office packages is included in the overall development cost of the software. There are however some circumstances in which end users or developers may be willing to pay for proofing tools, e.g.:

- if developers are keen to integrate text proofing tools into software to gain a pre-competitive advantage (e.g. where the software’s nearest rivals do not include proofing tools in their products)

[6] Included as standard for English. See http://www.microsoft.com/uk/office/cymruwales/default.mspx for a Welsh spellchecker for Office 2003; Welsh spellcheckers for other recent versions of Office are also available.
[7] See http://cy.openoffice.org/ or http://www.agored.com/
[8] http://www.languagetool.org/
[9] http://www.e-gymraeg.org/cysgliad/
[10] http://www.klebran.org.uk/

- where paid-for stand-alone products offer propositions to users that are not available elsewhere, e.g. Cysgliad, which retails for £55 includes the only comprehensive grammar checker available for Welsh, as well as other tools such as a suite of electronic dictionaries.

17.5. In spite of the wide availability of Welsh-language spellcheckers and/or grammar checkers for word processors and web browsers, many users appear to be unaware of their existence. There is also some confusion regarding the functionality provided by different proofing tools, with users failing to distinguish between products that are part of Microsoft Office and those which are third party add-ons. That Welsh texts often have a greater need for grammar checking rather than spell checking compared to English texts is also often overlooked.

17.6. For the reasons discussed in Sections 17.3 and 17.4, it is decided not to investigate text proofing tools further as part of this report.

17.7. Speech-enabled technologies encompass a wide variety of products, but the majority of them make use of speech synthesis, whether as a module or as an application in its own right. A full report on the state of the art in speech synthesis is found in Appendix F1. It finds that the field is reasonably advanced, to the extent that it is often difficult to impossible to distinguish between a synthesised voice and one 'recorded as live', especially if the vocabulary and grammar of the speech output is restricted.

17.8. Speech recognition is discussed in Appendix F2. In contrast to speech synthesis, speech recognition cannot always be guaranteed to work reliably, due to the probabilistic nature of the techniques involved, and the wide variation of human speech that even the simplest system has to recognise. However, the technology may have been said to have reached maturity in English and some other commonly spoken languages of the world.

17.9. Translation memory systems, also known as computer-assisted translation (CAT) systems, are now widely developed. The techniques underlying CAT tools are well-known and tractable, and CAT tools all operate within. Consequently, it is not believed that the underlying technical 'state of the art' in CAT software is likely to advance significantly in the near future. There are, however, significant differences in the functionality of available software, and this is described in Appendix F3.

17.10. In contrast to CAT, machine translation systems (which attempt to translate entirely automatically from one human language to another, with little or no human input in their operation) are the subject of active research. The underlying state of the art is in flux, to the extent that there are currently three different paradigms for machine translation systems, which are used either in combination or in competition with each other. The initial section of Appendix F4 (Machine Translation Paradigms) gives an overview of these classes, together with a brief explanation of the techniques underlying each one of them.

17.11. This section has given an overview of the state of the art in four of the technologies pinpointed as being of interest to SALT developers and users. The remaining technologies, from those listed in Section 17.2, are discussed in detail in Section 20 of this document.

18. Visits to key SALT laboratories

18.1. In previous sections, the state of the art in SALT has been mapped through surveys of available software and literature surveys of development techniques. These are useful in gaining general overviews of the relevant fields. However, it was also decided to visit laboratories renowned for the high quality of their SALT research. This was in order to gain the personal opinions of those individuals most heavily involved in leading-edge SALT development, and also to gain insights into developments more recent than those described in available literature.

18.2. In choosing the laboratories to visit, care was taken to reflect the wide-ranging nature of SALT as a discipline, as it encompasses applications both in speech and language processing. The categories outlined in Section 17.2 were also borne in mind, which reflect the interests of Welsh developers and users in fields as diverse as speech synthesis and machine translation. Therefore, it was decided to visit one researcher that mainly dealt with speech applications, and another whose interests were more concerned with written language.

18.3. The interview with Martin Russell of Birmingham University (Appendix G1) demonstrates the broad range of applications found even within a subdiscipline of SALT, namely speech technology. Martin Russell's work also demonstrates how SALT is growing in its scope to include multimodal technologies, i.e. applications which use speech as merely one of many modes of communication. This is seen by many as facilitating a more intuitive use of technology. Professor Russell argued that SALT should be part of a portfolio that reflects our normal everyday communication with other human beings, where we use gestures, movements and other non-verbal modes to enhance the understanding of the speech being uttered.

18.4. The Language Technologies Institute at Carnegie Mellon University, at which Lori Levin (Appendix G2) is an associate research professor, is world-renowned for its development of cutting-edge, widely used SALT components. Some of these, such as Festival (Appendix I1) and Sphinx (Appendix I2) will be discussed later in this document. The interview focuses on training in SALT, and how prospective university students can be encouraged to study the relevant technologies, thus building up the knowledge base. To this end, the Institute takes part in an annual competition for US high-school students, which takes the form of a Linguistics Olympiad. This is effective in building awareness and interest at secondary school level. The LTI also organizes postgraduate courses in language technology within Carnegie Mellon University, at masters and doctorate levels.

18.5. There are some key findings from these interviews that are pertinent to the SALT Cymru project, and which reinforce comments made during information gathering for previous sections of this document. These include:

- An emphasis on multimodality as a way forward for increasing the acceptance of SALT developments by end users. Martin Russell's emphasis on multimodality echoes the thoughts of Matt Jones on the subject (Section 12.3). Including multimodality in SALT developments requires generally that the SALT contained within them is sufficiently mature. With some exceptions, noted in Section 20, this is not the case for SALT in Wales. Therefore, multimodality will be kept in mind as an ideal for future developments, but will not be considered a short- or medium-term goal for SALT Cymru activity.

- The importance of adequate training in SALT, which reinforces Richard Sheppard's views on the subject (Section 12.5). High quality students should be nurtured from the earliest opportunity, and the model used by Carnegie Mellon's LTI, namely of organising competitions for school students who might be interested in language technologies to encourage them to study the subject in higher education. This is followed up by the university in undergraduate modules and postgraduate courses. Nurturing high quality students ensures not only the sustainability of SALT developments as they now exist, but also gives potential for the growth of the knowledge base in future.

19. Conference visit

19.1. The previous sections have given glimpses of individual views of the state of the art in speech and language technology. It is appreciated that an exercise of that nature, relying as it does on specific individual contributions, does not necessarily provide an adequate overview of the field at a specific point in time. Therefore it was decided to visit a conference offering a more comprehensive view of the state of the art in SALT, representing contributions from a significantly larger number of organizations and research institutes.

19.2. SALT is a wide field of research, and therefore it is rare to find conferences that provide a comprehensive overview of the field without concentrating solely on its speech or its language component. The LangTech conference, however, appeared to offer a view of both, and was thus attended by two of the project team.

19.3. A full report from the conference appears in Appendix H. While the conference was sufficiently wide-ranging to caution against drawing any specific conclusions from its activity, the following are general conclusions from the presentations:

- The main movements in SALT generally are to reduce costs and increase scales of services. This has implications in the public and private sectors. The private sector may use SALT, for example, to increase throughput in call centres. In the public sector, one specific example is of machine translation within the European Union government, which is being 'sold' to staff as a way of enabling gist translations into languages in contexts for which the Commission does not possess sufficient resources for human translations.
- The financial scale of SALT deployments internationally is significant and growing. The language industry generated $8.8bn in revenue in 2005. Spending on speech recognition globally was projected to be $3.2bn in 2010, a significant rise from $1.2bn in 2005.

19.4. In the Welsh context, interesting lessons can be drawn from some of the findings of the conference. SALT has to be of sufficient quality in lesser-resourced languages, to avoid creating 'noise' in languages that are less developmentally robust to inaccuracies. This is a challenge, as SALT will inevitably attract lesser levels of funding in lesser-resourced languages. However, technologies can scale up to make support for the majority of languages viable, and companies able to develop for smaller languages will reap rewards in multilingual world markets. This is due to lesser resourced languages providing a long tail of languages, as the number of speakers of lesser resourced languages is greater than the number of speakers of the top 30 languages of the world. Therefore, lesser resourced languages are a fertile area of research on not only scaling up technology but also in cross-disciplinary aspects related to the adoption of language technologies. As customer systems become increasingly self-service, and the scales of service that can be enabled through SALT increase, internationalization and localization of many aspects of SALT will be required in a continually greater range of languages and contexts.

20. Software packages in SALT

20.1. Five software packages are evaluated in this section. They cover the key areas of SALT

- Festival: speech synthesis
- Sphinx: speech recognition
- UIMA: the semantic web (intelligent web searching; keyword and trend spotting from text)
- Tesseract: optical character recognition
- Moses: machine translation

This section includes an overview of all the packages that have been evaluated, summarising the findings on each one, and highlighting priorities for future development.

20.2. Appendix I1 offers an overview of Festival, a popular open-source framework designed for speech synthesis development. It shows that its development for the Welsh context is at an intermediate level, offering functional voices that are well-integrated with popular operating systems. However, the voice quality for Welsh is currently lower than that for British or American English.

20.3. Appendix I2 discusses Sphinx, a set of open-source packages for speech recognition. In contrast to speech synthesis, speech recognition has not been significantly developed for Welsh, and any project will start from a very low level of development. However, this issue is to some extent being dealt with by a project carried out by the Language Technologies Unit at Bangor University, and funded by the Welsh Language Board for the financial year 2008/09, to develop basic speech synthesis for Welsh. While this will not result in anything approaching the level of maturity for English, it represents a foundation upon which future development can be built.

20.4. Appendix I3 investigates the UIMA framework, which offers a structured platform for information extraction. This provides significant possibilities for future development and high potential worth to Welsh SMEs, as it facilitates applications such as intelligent web searching, information retrieval and extraction. Unstructured information is the largest and the fastest growing source of information to business and organisations. Welsh SMEs could therefore be both developers and users of such technology. Any applications that could structure such information in a multilingual environment would provide savings and efficiencies for businesses beyond the original developers.

20.5. Appendix I4 discusses the Tesseract OCR (optical character recognition) system, which is available as a free, open-source standalone application for English, but due to its open nature, allows other languages to be included within its framework. There is currently no OCR technology that produces satisfactory results when scanning Welsh and bilingual Welsh/English printed text. The ability to accurately digitize Welsh language texts would benefit many sectors and expertise developed in the process could be put to commercial use in other languages.

20.6. Appendix F4 (in a separate section to the other appendices) discusses the Moses machine translation system, an open framework for developing the automatic translation by computer of one human language to another. It presents the results of a small pilot study in developing machine translation for the Welsh-English language pair. The results

are highly accurate for language close to the domain for which the system was originally trained, i.e. proceedings and legislation from the Welsh Assembly Government. It is less accurate for language divergent from this. This does, however, show promise for development of machine translation for Welsh and English and its use in certain restricted contexts, as long as sufficient bilingual textual data can be found for use in development.

20.7. In addition to the resources and potential resources for Welsh mentioned in the previous sections, there are other general aids to speech and language technology that have a role in future development in the field. Examples include corpora of speech and language, required for applications such as speech recognition and machine translation. Some work has already been undertaken in these fields. The SpeechDat Welsh database, collected between 1996-99, presented digitized recordings of 2000 Welsh speakers over telephone lines. It was designed for the development of voice-driven telephony services. A written electronic corpus, CEG, consists of a million words of Welsh, and has been used widely, particularly in the development of spelling and grammar checkers for the language, and in speech synthesis applications. Welsh still lags behind more widely-spoken languages in language resources of the kind, and there is a clear need both for a larger written corpus and a non-telephone speech corpus. Both of these are seen as essential aids for the development of SALT in any language.

20.8. It is noted that all the packages described in this system are open-source ones. This is a deliberate decision. Such packages can be developed in a non-state aid environment such as the ones that exist for many funding streams exploited by Welsh academia and industry. A successful model for this, used by the Language Technologies Unit, Bangor University, in previous projects, has been to develop tools under a liberal free software license. This license allows the tools to be used in other projects, whether commercial or free, without restriction. Therefore, the tools can be developed through grant aid and versions of them released to the public at no cost. Outside the scope of these projects, other versions can be developed from the free ones, either by academia under exclusive or non-exclusive contracts to businesses, or by businesses possessing sufficient knowledge to develop the products themselves. Such non-free products can be tailored to businesses' individual needs, or can provide refinements not present in the free versions. Businesses can then resell or further develop these versions, offering them to end-users at cost.

21. Analysis of coordination methods

21.1. Many methods have been used in the course of the project to attempt to bring those interested in SALT together. The relative successes of these are discussed below.

21.2. The main method of gaining names for the project was through web signup. An encouraging amount of participants (over 70 names) were gathered in this way. It was found, as expected, that certain industry sectors predominated in the signup, which can be explained by word of mouth within sectors resulting in increased interest within them.

21.3. All the 70 participants were sent personal invitations to fill in the SALT Cymru survey. The resulting uptake was encouraging: 48 complete forms were received. The survey was carried out electronically, to increase uptake and simplify the process of completing the form. The survey was encouraging not only in the number of respondents but also in their range of interests. Respondents included users, developers and potential developers in SALT.

21.4. Focus groups showed mixed success. It was originally anticipated at project commencement that two groups would be convened as part of the project. The first focus group was successful. Comprising over 80 attendees, it had been organised as part of an already existing meeting (the annual conference of the Association of Welsh Translators). By contrast, another focus group could not be convened. This was not arranged as part of any existing meeting, and there were consequent problems in scheduling mutually convenient times for participants.

21.5. As a result of the above findings, two key recommendations for SALT Cymru specialist interest group activity are made. The first is that in convening group activity, use should be made if at all possible of already existing meetings, which may be augmented by group dissemination activities. This provides a means to break out of the vicious circle of a lack of initial awareness/interest in SALT leading to individuals placing low priority on attending any events that seek to encourage its use and uptake.

21.6. The other key recommendation is that use should be made of existing networks, whether formal or informal, to spread awareness of the specialist interest group. Individual businesses in similar sectors share a characteristic with speakers of lesser-resourced languages, in that neither exist in a vacuum but have contact with others of similar interests. These networks are not necessarily geographically based ones. Hence, spreading news of the specialist interest groups amongst these networks should encourage greater uptake and mutual enthusiasm.

22. Findings

22.1. A major finding of SALT Cymru, and one that raises concerns regarding the future stability of SALT development in Wales, is the fragility of the knowledge base. In the academic sector, fewer than ten full-time employees work within Wales on SALT. Of these, only about half are permanently contracted to do so. The situation in industry is also weak – the results of the survey found that all SMEs undertaking SALT activity also developed other technologies. It is likely that the reason for this diversification is that the SALT sector in Wales is still at an early stage of development, making it difficult to sustain businesses that develop SALT alone. However, it is encouraging to find that academia and industry are already engaging together in knowledge transfer activities. The current KTP in the field of speech recognition and machine translation with Testun Cyf. (see Appendix C5) is an excellent example of how a Welsh SME can benefit from SALT through partnering with a HEI in Wales. It is also an example of how catering first for the home market is expected to open export markets in the multilingual international community.

22.2. Coupled and associated with the fragility of the knowledge base is the low level of awareness amongst potential users (and some potential developers) regarding what SALT has to offer them. This is to some extent understandable; SALT is a fast-developing discipline and many SALT techniques have only reached maturity comparatively recently. However, there are other scenarios where technologies that are already mature are not used to their full potential. One typical area highlighted both by the survey and by the interview with Tegau Andrews (Section 12.5) is the low uptake of translation memory systems within the Welsh translation industry. This impacts not only the translation industry itself, but also the customers of that industry, such as local authorities in Wales. Given the potential for time and cost savings, it is surprising that more use of CAT tools is not made in Wales and that the public sector in particular are not heavy users of the technology.

22.3. It was found, especially in the focus group of translators, that the appetite for SALT solutions increased as the offerings were explained. Clear explanations were given of what various parts of SALT entailed, and discussions included the application of SALT to everyday issues faced by translators. The level of interest shown in SALT following the discussions was significantly higher than previously. Although bilingual administrators in the public sector were not directly approached for their views, indirect evidence, again from the research of Tegau Andrews (Appendix C3) and the interview with Richard Sheppard (Appendix C4), suggest that there is further room for clear explanation of the benefits of SALT solutions to those in charge of bilingual administration in Wales.

22.4. The acronym SALT has been used in industrial and academic circles for many years to describe Speech and Language Technologies. However, it did cause some confusion in the project survey as it is also in active use in Wales as an acronym for Speech and Language *Therapy*, the medical discipline concerned with therapy for speech and language disorders. During the lifetime of the SALT Cymru project, many speech and language therapists signed up for the project, despite clear explanations on the project website that the focus of the project was on technology. Some of these therapists were later disappointed upon realising that the project aims were at best peripheral to their occupational activities. However, given that the SALT Cymru ‘brand’ is now reasonably well-known as a result of the project’s activities, it is not felt wise to change its name.

22.5. Users' expectations of SALT are high, especially if they are marketed (as most are) as productivity aids for businesses and individuals. Care needs therefore to be taken to manage expectations. Due to the wide-ranging nature of human language, SALT applications cannot be guaranteed to work reliably in all contexts. It is noted that one company, operating outside Wales, is currently marketing an online translation system between English and Welsh, also available as a standalone product for a PC, which is regarded by professionals as very low quality. No disclaimers appear on the company's promotional or online material regarding possible inaccuracies in the translations. As a result, some highly erroneous, almost nonsensical, translations have appeared on public notices: these have subsequently had to be removed and professionally translated[11]. In fields such as speech recognition, the day-to-day variability of speech from the same speaker may even mean that the results may not even be consistent from one hour to the next. In such cases software licenses normally contain clauses disclaiming the nature of the statistical processes underlying the technology. This is an example of good practice in SALT, and all members of any specialist interest group should be encouraged to follow such practices.

22.6. It is believed that the majority of SALT developers are reputable, and concerned above all about the quality of their product. However, as stated in Section 22.5, there are some SALT tools which are not fit for purpose, or which are marketed in a way which perhaps overstates the quality of results that may be achieved by using the tool. Educating users is therefore as important as educating developers.

22.7. Sustainability is a key theme in much economic development work. In environmental terms, SALT has a positive effect on sustainability: there are efficiencies gained in running SALT modules which have a small positive effect on the energy consumption of those using them. These are likely to become more important in future years as environmental issues take centre stage.

22.8. Taking sustainability in the broader terms of the sustainability of SALT development within Wales, achieving sustainability in SALT is more likely if the components developed can be reused widely across applications and languages. It is intuitive that the probability of reuse increases if SALT components are developed with multilinguality in mind. One of the core philosophies of the SALT Cymru project is to regard the bilingual Wales as a living laboratory for the development of multilingual techniques. Such multilingual techniques thus open the doors to global markets, making the sustainability of research and knowledge more attainable.

22.9. Some open source utilities developed for cross-lingual use such as those described in Appendix I of this document may be further developed as basic resources for a language toolkit for SALT developers in Wales. The adaptability of such utilities for use with multiple languages points a way for the needs of less-resourced languages was confirmed. This provides a sustainable method of building basic language resources both for end-users and for further development by SMEs.

22.10. Drawing up recommendations for the establishment of a SALT Cymru specialist interest group was one of the stated aims of the present project from the outset. Interest in the project and responses to the questionnaire indicate that this is both viable and

[11] See, e.g., http://news.bbc.co.uk/1/hi/wales/5341646.stm

desirable. Innovative ideas, such as the establishment of a Language Resource Portal (see Appendix C4), came to light as a result of the consultation exercise. Other solutions to the specific circumstances of industry and academia in Wales are likely to emerge from such a group, to the benefit of both. Better communication and liaison between interested parties will solve the present problems of fragility, fragmentation and low take-up. A specialist interest group will be well placed to implement other recommendations detailed in the following section (section 23).

23 Recommendations for future action

23.1. Establishment of a SALT Cymru specialist interest group (SIG).

The 44 positive respondents to the SALT questionnaire should form the initial core of this group. However, open invitations should be given to other interested parties in Wales to join in order to maximize effective communication within the sector. The group should contain separate strands for developers and users of SALT in a flexible arrangement that ensures cross-fertilization of ideas, joint projects and training opportunities.

23.2. International Watching Brief

The SIG should maintain a watching brief on international developments to convey forthcoming trends, innovations and opportunities to developers and users in Wales. It should also ensure that information is a two way process, promoting the activities of SALT developers in Wales in the global market and raising the profile of Welsh academia and industry in the international multilingual context. It should use such opportunities as that afforded to it by the hosting of the GUM3C (Global Understanding in Multilingual, Multimodal and Multimedia Contexts) Conference by the Language Technologies Unit, Bangor University, in August 2008 (see http://www.gum3c.org/) to further this aim and engage with the international academic and industrial community.

23.3. Strengthening the Research Base.

The SALT research base in Wales should be consolidated with its international level standing and participation strengthened. The SIG should influence and guide the research base in its efforts at enabling and coordinating funding applications for new SALT projects to research bodies in the UK and further afield. The emphasis should be on joint projects and world class centres of excellence with a view to promoting joint cooperation between HEIs in Wales and between HEIs and industry. Due regard should also be given to seizing joint international opportunities, such as those afforded by EU projects such as FP7 and Interreg IV Wales/Ireland programme, to strengthen cross border SALT research.

23.4. SALT Cymru Website and Resource Portal

The present SALT Cymru website (www.saltcymru.org) should be maintained and extended as a one stop shop to provide information and resources for SALT developers and users in Wales. It should include provisions for virtual meetings, an on-line newsletter, details of forthcoming workshops, conferences and other events of interest to its members. It should also include a Resource Portal where the language resources developed in 23.5 could be posted.

23.5. Developing a Basic Language Resource Kit.

A prioritized programme for basic language resources should be established by the SIG with useful utilities being made publically available for download to software developers in Wales. Any existing openly available software, including those evaluated by the SALT Cymru projects, should be included, adapted and developed further. Investment and

funding should be attracted to ensure sustainability and quality control issues considerations.

23.6. Evaluation and Quality Control

The SIG should consider the appropriateness of introducing evaluation of SALT components produced in Wales. This could include an accreditation scheme and/or identifying relevant international industry techniques, practices and/or standards, and providing training in their deployment.

23.7. SALT Training and Education

A comprehensive programme of training and up-skilling in SALT issues should be undertaken, with the aid of the SIG. There should be a clear path for SALT developers from school through undergraduate to postgraduate level and into industry. Rather than create new courses, the emphasis should be on identifying clear paths to use existing resources, creating new modules if needed, and having due regard to the multidisciplinary aspects of the subject. Negative comments about the quality of the present provision should be taken on-board, with a view to providing training paths that are fit for purpose. Successful programmes such as KTP should be used as much as possible to transfer knowledge between academia and industry in the SALT sector.

23.8. Up-skilling the Workforce and Improving Communication

The SIG should address the training needs of its own members, especially of SALT users and developers of products other than SALT utilities for who the incorporation of SALT components in their output who have themselves no expertise in SALT. Where possible, the SIG should use existing events to promote its activities, e.g. seminars and conferences already planned for the Translation Industry, rather than organise its own events in competition. It should however plan for at least an annual conference, with regular seminar programme, organising such events itself if it is unable to identify suitable pre-existing activities.

23.9. Funding

The recommendations outlined above seek to make frugal use of resources, with due regard to sustainability. Funding opportunities exist from sources such as UK Research Councils and the EU. However, the establishment of a sustainable SIG and the activities outlined above do not fall neatly into the remit of any of the above. Given the magnitude of opportunities for the Welsh research base and economy along with the benefits to internal bilingual and multilingual communication, the Welsh Assembly Government should be encouraged to provide the funding necessary to place the SIG and the activities recommended above on a secure footing at its inception.

24. Conclusion

It is believed that this document represents a workable framework for the development and the encouragement of uptake of speech and language technologies within Wales. It provides opportunities to kick-start the latent demand and interest in SALT, in a manner that will greatly benefit not only the productivity of users of the technology, but also the profitability of current and future SALT developers in Welsh industry.

Appendix A: Definitions of SALT from various sources

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA

- **Main website:** http://www.lti.cs.cmu.edu/
- **Permanent link to the definition:** http://www.lti.cs.cmu.edu/
- **Definition:** *The Language Technologies Institute (LTI) at Carnegie Mellon University (CMU) conducts extensive research on Computational Linguistics, Machine Translation, Speech Recognition and Synthesis, Information Retrieval, Computational Biology, Machine Learning, Text Mining, Data Mining, Knowledge Representation, and Intelligent Language Tutoring. Our "Bill of Rights" is: Get the right information (search engines, question answering, text mining) to the right people (adaptive filtering, personalization) at the right time (task modelling, anticipatory analysis) in the right language (machine translation, cross-lingual retrieval) and the right media (speech recognition and synthesis) at the right level of detail. (summarization, question-answering, drill-down)*
- **Notes:**
    - This definition does not seek to comprehensively define SALT, but rather lists areas within the field in which research is undertaken by the Institute.

Language Technology Group, Melbourne University, Australia

- **Main website:** http://www.cs.mu.oz.au/research/lt/
- **Permanent link to the definition:** http://www.cs.mu.oz.au/research/lt/
- **Definition:** *Most human knowledge, and most human communication, are represented and expressed using language, both in written and spoken forms. Language technologies permit computers to process human language, providing more natural human-machine interfaces, and more sophisticated access to stored information. Language technologies will play a central role in the multilingual information society of the future.*
- **Notes:**
    - Steven Bird, a member of the Language Technology Group, was a co-director of the Linguistic Data Consortium (http://www.ldc.upenn.edu/) before moving to Melbourne. Emphasis, therefore, has been placed both on the written and spoken forms of language in this definition, given that the LDC dealt with the collection of written and spoken speech and language corpuses for SALT research.

Language Technology Group, Edinburgh University

- **Main website:** http://www.ltg.ed.ac.uk/
- **Permanent link to the definition:** http://www.ltg.ed.ac.uk/
- **Definition:** *[The LTG] focus[es] on building practical solutions to real problems in text processing. We have worked in all areas of large-volume text handling, from text annotation through markup architectures and from information extraction to automatic or computer-assisted generation of text.*
- **Notes:**
    - This is not a comprehensive definition of SALT. It is, however, worth noting the way in which it reflects the fragmentary nature of SALT research, even within many universities. The Language Technology

Group in Edinburgh is part of the Human Communication Research Centre (http://www.hcrc.ed.ac.uk/) . The University also contains the Centre for Speech Technology Research (http://www.cstr.ed.ac.uk/) but this is a separate research centre
    - The lack of emphasis within the definition on speech technology, therefore, reflects a lack of emphasis on speech technology within the *department* in Edinburgh, but not within Edinburgh University itself.

The Edinburgh-Stanford Link: Language Technology

- **Main website:** http://www.edinburghstanfordlink.org/
- **Permanent link to the definition:** http://www.edinburghstanfordlink.org/lang_intro.html
- **Definition:** *Language technology refers to a very broad range of human computer interaction (HCI) technologies that have been developed over the last 20 years to enable people to more easily and naturally communicate with computers, through speech, text or gesture, and when called for, receive an intelligent and natural reply in much the same way as a person might respond.*
- **Notes:**
    - This is a clear and concise definition, which while not giving a precise definition of those technologies which do and do not constitute SALT, is nevertheless very suitable for a lay audience (see also the remainder of the page http://www.edinburghstanfordlink.org/lang_intro.html)
    - It could be argued, however, that 'the last 20 years' is too specific a phrase. Speech and language technologies have existed in one form or another almost since the advent of the digital computer in the late 1940s.

Hans Uszkoreit, Saarland University: What is Computational Linguistics?

- **Main website:** http://www.coli.uni-saarland.de/
- **Permanent link to the definition** http://www.coli.uni-saarland.de/~hansu/what_is_cl.html
- **Definition:** *Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty. It belongs to the cognitive sciences and overlaps with the field of artificial intelligence (AI), a branch of computer science aiming at computational models of human cognition. Computational linguistics has applied and theoretical components.*
  Further: *Applied CL focuses on the practical outcome of modelling human language use. The methods, techniques, tools and applications in this area are often subsumed under the term language engineering or (human)* ***language technology.*** [our emphasis]. *Although existing CL systems are far from achieving human ability, they have numerous possible applications. The goal is to create software products that have some knowledge of human language. Such products are going to change our lives. They are urgently needed for improving human-machine interaction since the main obstacle in the interaction between human and computer is a communication problem. Today's computers do not understand our language but computer languages are difficult to learn and do not correspond to the structure of human thought. Even if the language the machine understands and its domain of discourse are very restricted, the use of human*

language can increase the acceptance of software and the productivity of its users.

- **Notes:**
  - The original SALT Cymru bid referred to SALT as the *engineering branch to computational linguistics*, therefore the definition of *computational linguistics* is of key importance in defining SALT itself.
  - While the above definition restricts itself to language technology, the following definition by the same author discusses speech technology in the SALT context: http://www.dfki.de/~hansu/LT.pdf

Appendix B: Departments and units in Welsh HEIs undertaking SALT research and SALT-associated research
(Draft report – this is not an exhaustive list)

Bangor University

http://www.bangor.ac.uk/

- Canolfan Bedwyr's Language Technologies Unit has a team of eight staff (equivalent to five full-time staff) undertaking SALT research in a variety of fields
- The ESRC Research Centre for Bilingualism in Theory and Practice undergoes a number of SALT-associated research projects, but is not a SALT developer in and of itself
- Dr William Teahan in the Computer Science department (http://www.cs.bangor.ac.uk/Staff/william_teahan.php) has a variety of research interests in SALT:
    - Classifiers and Text/Data Mining is a SALT research field: http://www.cs.bangor.ac.uk/research/text_data_mining.php
    - Artificial Intelligence and Intelligent Agents is SALT-associated: http://www.cs.bangor.ac.uk/research/aiip.php

Swansea University

http://www.swan.ac.uk/

- The Speech and Image Research Group in Electrical Engineering undertakes some SALT projects, including speaker recognition work and related biometric research, and has in the part conducted research into Welsh-language speech recognition. The head of the group is Dr John Mason http://galilee.swan.ac.uk/
- The Centre for Applied Linguistic Studies (CALS) notes various interesting projects related to SALT, including the development of computer software to assess the vocabulary of learners of English as a second or foreign language. Professor Paul Meara is the relevant person here. http://www.swan.ac.uk/cals/ is their old website.
- Future Interaction Technologies is a centre within Computer Science. It conducts research on making computer interfaces (including mobile phone interfaces) more user-friendly. This is multimodal research rather than SALT research per se, but may well impinge on SALT-related fields in the future. SALT is seen by them as a modality in the context of their research. One of the centre's members (Dr Matt Jones) has a PhD in the field of speech recognition. Their website can be found at http://www.fitlab.eu/, and an interview with Matt Jones can be found in Appendix C of this report.

Glamorgan University

http://www.glam.ac.uk/

- SALT-associated research is conducted by Daniel Cunliffe, who investigates the use made by minority language speakers of various language technologies. He is not a pure SALT developer in that sense, but the results of his work can

illuminate future research in SALT fields. His web page can be found at http://hypermedia.research.glam.ac.uk/caml/ and his weblog can be found at (http://datblogu.weblog.glam.ac.uk/). He is interviewed in Appendix C of this report.

University of Wales Lampeter

http://www.lamp.ac.uk/

- Some SALT research in online language tools is conducted by the Welsh Department. In particular, http://www.geiriadur.net/ is a popular online Welsh-English

Aberystwyth University

http://www.aber.ac.uk/

- The Mercator Centre http://www.aber.ac.uk/~merwww/ conducts research regarding, and has a large number of links to, the film and television industry in Wales including a significant number of independent companies. This is not SALT research as such, but they are likely to be interested in any SALT Cymru outputs given the Mercator centres' links to those researching in other lesser-spoken languages in Europe.
- The University has a visualisation group, which while not conducting SALT research, is likely to be interested in the outputs of SALT Cymru: http://www.aber.ac.uk/compsci/public/research/research-groups/vision-graphics-and-visualisation.php

Cardiff University

http://www.cardiff.ac.uk/

- The Department of Welsh has several projects to create electronic (digitised) versions of old Welsh texts. This is not direct SALT research, but makes use of SALT outputs (hence is SALT-associated research). Work by them on a digitised corpus of Welsh ballads was undertaken in conjunction with Canolfan Bedwyr's Language Technologies Unit. http://www.cardiff.ac.uk/cymraeg/english/research/researchProject.shtml
- Centre for Language and Communication Research (SALT-associated): http://www.cardiff.ac.uk/encap/clcr/resact.html
- Communication Research Centre. SALT-associated – work from the psychological angle researching the man-machine interface. http://www.cf.ac.uk/psych/crc/index.html
- Centre for Digital Signal Processing. Though this is SALT-associated (digital signal processing techniques are used extensively in low-level research in SALT) the Centre does not appear to conduct SALT research as such. http://www.cardiff.ac.uk/engin/research/informationsystems/cdsp/index.html

Appendix C: Interviews with key players in SALT

C1: Interview with Dr Matt Jones, 5th December 2007

The interview was conducted in Swansea University, where Matt Jones is a Reader in the Future Interaction Technologies Lab (FIT Lab).

About the FIT Lab

The Lab (http://www.fitlab.eu/) was formed in 2005-6, to bring together existing researchers in the field in a centre of excellence based at Swansea University. The senior personnel in the lab are as follows:

- Prof Harold Thimbleby (Lab Director, formerly of UCL)
- Dr Matt Jones (Reader, formerly of Waikato University, New Zealand)
- Dr George Buchanan (Lecturer)
- Dr Parisa Eslambolchilar (Lecturer)

At the time of writing (January 2008) there are two research assistants in the group, with a further two currently being appointed. There are five PhD researchers, one of whom also fulfils a research assistant role. An additional PhD researcher is currently being appointed. There is also one MRes researcher.

It is noted (in another interview with Matt Jones at http://www.itwales.com/997744.htm) that the FIT Lab MSc started in October 2006, with some funded places from ITWales.

Research Projects

The main focus of the FIT Lab is *usability* of computer systems in all their forms. The Lab's particular research focus is on user interface (UI) design, and Matt Jones' focus is on mobile devices (book: *Mobile Interaction Design*, Matt Jones and Gary Marsden, Wiley, 2006). The Lab undertakes some projects with and for Nokia Research Centre in Helsinki.

I was shown a prototype mobile device which comprised an internet-connected Pocket PC with a GPS add-on. The concept was to use mapping information available from the Internet and combine it with the GPS location information, thus presenting the user with a live map of their area.

One of the Lab's other projects also involves novel GPS devices, in the shape of an MP3 player which integrates GPS information to guide users to their destination. As they walk directly to their target, the user's chosen music is played at full volume in both ears of their headphones. As they veer away from their target, their music is panned to the left or the right ear, and increased and reduced in volume, thus giving them live guidance on whether they need to change direction or not. Both this and the live mapping device are being further developed through the EPSRC-funded *Multimodal Negotiated Interaction in Mobile Scenarios*, the principal investigator for which is Professor Roderick Murray-Smith at Glasgow University (see http://www.dcs.gla.ac.uk/~rod/)

The Lab is heavily involved in digital storytelling projects, and their StoryBank (http://www.cs.swan.ac.uk/storybank/) project links schools in Swansea with those in India. A rugged laptop is used to record children's and other community stories in Wales and India. They are edited by the communities themselves with assistance from Lab members, who then visit both communities to play back the other's stories. The research in StoryBank centres on the playback interface between the user and the stories themselves. Rather than force the user to interact through the keyboard or mouse, they touch numbers on a screen to play back stories with that particular ID. The StoryBank project is funded by the EPSRC's Digital Divide programme.

FIT Lab and their relationship to speech and language technology

It became apparent in conversation that speech and language technologies are seen by the FIT Lab as a means to an end, rather than an end in themselves. The FIT Lab deals in multimodal interfaces. Multimodality, in computer science, can comprise anything which moves human interaction with computers beyond the now traditional keyboard/mouse interfaces. Therefore, speech input and output would be seen as one possible mode of human-computer interaction amongst many others.

It is tempting therefore, in the terminology of SALT Cymru, to see the FIT Lab as *users*, rather than *developers* of SALT. In their case, however, this is not a sharp distinction: the integration of SALT into multimodal interfaces is certainly part of their work, which constitutes integration, and therefore some de facto development of the technologies themselves. It is also relevant to note that Matt Jones has a speech technology background: his PhD (Cambridge, 1994) examined acoustic modelling for large vocabulary continuous speech recognition. In the past, he has contributed to the widely used HTK speech recognition toolkit from Cambridge University's Engineering Department, and to the original Microsoft Speech Application Programming Interface, developed by Entropic, a company spun out of the department.

C2: Interview with Dr Daniel Cunliffe, 12th December 2007

The interview was conducted in Glamorgan University, where Daniel Cunliffe is a Senior Lecturer in Multimedia Computing, in the Division of Computer Science.

Daniel Cunliffe's research touches on aspects of SALT, particularly language (rather than speech) technology. However, he is principally interested in the *use* made of the technology by people, rather than the technology itself.

Research areas of interest

Some questions asked by Daniel Cunliffe's research include:

- On websites and in other IT situations where more than one language is available to the user, what influences that choice of language?
- How can Welsh be made available on every computer in Wales, and further, how can users be encouraged to use Welsh on their computers?
- How can language planning and language policy influence the uptake and use of IT in the Welsh language?
- Should computing environments reflect the language choice of the user, or try to change it?

Research projects

One interesting research project carried out by Daniel Cunliffe and his research assistant, Andrew Deere, into the use and effectiveness of 'Welsh Google'. This is merely a Welsh-language translation of Google, and does not include all the capabilities one might expect from a Welsh localisation of a search engine. For instance, it does not include the ability to pick up all forms of words, mutated and unmutated, when the user enters a search query.

The work done on investigating 'Welsh Google' raises an important point in the context of SALT, namely that difference between those who are content to work around the limitations of such technology (in the case of Google, maybe by searching for mutated forms of words in Welsh) and those who may not even be aware that such limitations exist.

Jointly with Courtney Honeycutt, who is working towards a PhD in Indiana University's Bloomington School of Library & Information Science, Daniel Cunliffe is researching the use of weblogs (blogs) in the Welsh language. This is being achieved by a study of blog posts and comments over a three-month period, including an investigation of content, linking and community.

Daniel Cunliffe is also researching the use of language on party political websites during the Welsh Assembly elections in 2007.

Groups of interest

Daniel Cunliffe is involved with the setting up of the British Computer Society's South Welsh branch (see http://southwales.bcs.org/)

Requirements in research

I asked Daniel Cunliffe whether he had a 'wish-list' of items that would aid his research. Some points raised in the resulting discussion included:

- A way of bringing together publications regarding the Welsh language and technology (this wish highlights the multidisciplinary nature of fields related to SALT)
- An easy way to find and tap into a group of Welsh speakers to investigate their use of Welsh-language technology. It is not easy to find external funding for such initiatives. In order to give a realistic overview of the actual use of technology in the population at large, this group should not exclusively contain very active users of Welsh in technological contexts (i.e. should not confine itself to the likes of Welsh-language bloggers). Questions to be asked to this group would include:
    - What is the actual language use of these speakers on their computers?
    - Why do they use language in the way they do?

Daniel Cunliffe and SALT

It is evident from the above discussion that Daniel Cunliffe is not a pure SALT developer. However, he does research the effects of SALT and its related fields, and the output of his research has, potentially, an important part to play through informing of the use made of SALT. This can enlighten us on why, in some cases, there exists a gap between the expected use of such technologies and the actual uses made of them.

C3. Interview with Tegau Andrews, 28th March 2008

Tegau Andrews is a PhD student researching the translation industry in Wales, with particular regard to the translation of web sites and the use of language technology tools within the industry. The PhD is sponsored by Ffilmiau'r Nant, under the Objective 1 funding programme for PhDs at Bangor University. When completed, it will be a very valuable contribution to the understanding of this important sector of the Welsh economy. In the meantime, this interview was conducted in order to gain a preview of the research results, and an insight into issues of relevance to the SALT Cymru project.

As part of the PhD research, Tegau Andrews conducted an extensive questionnaire with professional translators in Wales. This included translators working in both the private and the public sectors, those working in large offices and teams, and freelance translators working on their own. A key question asked of all translators concerned their use of computer-aided translation (CAT) tools in their work. These include, in particular, the use of translation memory (TM) systems, and tools facilitating the translation of web pages held in HTML or XML format. These tools are regarded as essential translation aids in the international translation industry, aiding productivity, quality control and management of documents and projects. For the purposes of SALT Cymru they may be regarded as a specialized category of language technology tools, facilitating translation, with additional features to integrate dictionaries, extract terms, create bilingual lexicons and even manage the translation process.

One of the key findings of Tegau Andrews' questionnaire was the low uptake of these CAT tools in Wales, with only 47% of respondents reporting their use. This varied between different types of translation agencies. They tended to be used more by private translation companies than by freelance translators. Surprisingly, there was a low uptake by local government authorities, despite their having in-house translation services, extensive translation needs to their statutory bilingual policies, and large bilingual websites.

The lack of use of CAT tools in the translation of websites was particular cause for concern. Websites often need to be continuously changed and updated, and if CAT tools are not used for this task, manual retranslation may be needlessly slow and costly. An additional concern was the technical quality of the work. If a translator is given the text as a HTML or XML file without software to ensure that the computer code is not disrupted or accidentally changed, grave errors may occur, and the data may not display correctly on the published web page. This is a serious matter on a bilingual website, and even more so in a multilingual website, such as the projected Ffilmiau’r Nant site which was to contain five languages, where such problems would be compounded according to the number of languages.

The lack of take-up of these tools in Wales may be attributed to a number of factors. Although there was a general awareness of the existence of such tools, there seemed to be little appreciation of the scale of savings that could be made through their use, both in terms of time and money.

A number of translators had no training in the use of these tools. Some providers offered training in the use of CAT tools as part of their marketing strategy, but this training was tied to particular proprietary software, and was not widely taken up. Some workshops were also provided by industry associations such as Cymdeithas Cyfieithwyr Cymru.

Not only was there a need for additional training in the use of CAT tools, both at a basic and advanced level, but there was also a need for more awareness of CAT tools outside of the translation industry, among web designers and other professionals who interacted in the process of producing and maintaining bilingual and multilingual documents and websites.

Through Tegau Andrews' work on the Ffilmiau'r Nant website, and her discussions with companies such as Microsoft and Lionbridge, it was seen that the decision to create systems which facilitate easy and fast updating of bilingual or multilingual websites rested with webmasters and administrators, not with translators. In cases where these personnel are unaware of CAT or any other SALT tools which could aid their work, this can greatly add to the difficulties of localising a website. Lack of CAT tool awareness may also create problems elsewhere - for example, when the managers of in-house translation teams are not aware of the benefits of such tools, the entire team may suffer. This may have been the reason for the low uptake of CAT tools by local authorities in Wales.

The position with freelance translators was somewhat different. Although they had more freedom to buy their own CAT tools, customers gave them no guidance on dealing with website formats, which left them in the dark. They needed advice but did not know where to find it, and they also felt that their training needs were unmet. Freelance translators are often used by large translation agencies, and the point was made that international companies would not deal with freelancers who were unable or unwilling to use available technology.

Lionbridge, one of the world's largest localisation companies, had an office base in Dublin, the Republic of Ireland. From here they distributed translation work to translators across Europe. They are considering establishing an in-house office for Irish language translation work. This is due to the increase in the demand for Irish language translators since Irish was made a full official language of the European Union in 2006, and due to the current lack of skilled Irish language translators. Lionbridge and its outsourced translators depend heavily on CAT tools. Some at Lionbridge drew attention to the fact that universities provided little relevant training for trainee translators, and that they lacked engagement with real world translation issues such as the use of CAT tools.

The Welsh language translation industry is more mature than its Irish language counterpart. In Wales, Cymdeithas Cyfieithwyr Cymru, with its office in Bangor, is a professional association which has an accreditation scheme for translators and has provided networking opportunities for translators for a number of years. There is no comparable network of Irish translators. However Foras na Gaeilge (the governing body of the Irish language) has established an exam-based accreditation scheme. Bangor University also provides an active network for translators through its welsh-termau-cymraeg discussion list. Both Cymdeithas Cyfieithwyr Cymru and Bangor University provide some training in CAT tools for translators, sometimes in cooperation with each other. Given the synergies between the Welsh and Irish translation industries, and the close proximity of the two regions, there were further possibilities here for cooperation, especially in improving the skills of translators and other professionals and providing training in the use of CAT tools.

We await the completion of Tegau Andrews' PhD with interest, and thank her for this opportunity to preview its findings for the SALT Cymru report.

Appendix C4: Interview with Richard Sheppard, Managing Director, Draig Technology Ltd, 26 March 2008 (with additional e-mail correspondence)

Background

Draig Technology is a software development company with offices in both Bangor and Cardiff. It offers both bespoke software solutions and delivery of software and project management consultancy services and prides itself on its reputation as Wales' leading software house. Although it does not market itself as a company engaged in SALT activities, many of its products and services have strong bilingual and multilingual capabilities and components. Because of its expertise in bilingual software issues it was commissioned by the Welsh Language Board to write the Bilingual Software Guidelines and Standards[12], as part of the Welsh Assembly Government's strategy of the Welsh Language, as expressed in its strategic document, *Iaith Pawb*[13].

Numerous case studies on Draig Technology's website[14] point to the importance of aspects of SALT in their work. For example, on the question of Language Support, Draig states that "there are several aspects to how an IT system, specifically the software can support a language. General tools such as spell checkers and thesaurus are important, however, the user interface and the management of data in multiple languages is also critical, yet is often overlooked." Furthermore, they define the nature of the language support required in IT applications to include:

- "Allowing the user to select their preferred language;
- Sensing that preference where it is not explicitly stated;
- Text rendering and management tools with consistency and availability checking;
- Diacritic mark support (i.e. ô, á, ŵ), allowing them to be displayed, stored and entered by the user;
- Tolerating diacritic equivalence (i.e. use of o in place of ô) in functionality such as text searches;
- Support for non-Latin alphabets, including digraph letters (i.e. Ll, Dd, Ng) and differing sort orders;
- The mapping of standard data (counties, months, etc.) to a language-neutral internal representation to ensure equivalent interpretation across languages;
- Multi-lingual data management to maintain the integrity of data stored in multiple languages;
- Compliance with language specific standards and guidelines, particularly e-government, accessibility and legislative."

Amongst the companies Draig has worked with to provide bilingual and multilingual software solutions have been Microsoft® (technology partner to localize Windows XP and Office 2003 into Welsh); South Wales Fire and Rescue Service (bilingual Content Management System as part of a broader package of support); and Cynnal Cymru

[12] Published in 2006 on the Welsh Language Board website http://www.bwrdd-yr-iaith.org.uk/cynnwys.php?cID=6&pID=109&nID=2063&langID=2

[13] Published in 2003 by the Welsh Assembly Government, see http://new.wales.gov.uk/topics/welshlanguage/iaithpawb/?lang=en

[14] See http://www.draig.co.uk/casestudies/?lang=en&plain=false

(Provision Management Database including multilingual support, multilingual user interface, language preference tracking and integrity reports).

Potential for growth for SMEs in Wales using SALT
Richard Sheppard was asked whether, in his opinion, that there was potential for growth for SME developers of SALT in Wales. There were two ways that this could happen:

"The first being for bilingualism within Wales and the second for a multilingual capability for broader/global markets.

For the first, the legislative requirement under the Welsh Language Act and organisations obligations under their Welsh Language Schemes that require them to provide bilingual communications/capabilities is a market driver. However, we have found the market to be extremely limited with very few organisations paying anything more than lip service to this. Even when there is a negligible cost, most public sector organisations are dismissive - maybe not publicly, but at the individual decision maker this is certainly the case.

The private sector is very much the same, though there is more honesty here in that a simple demographic and cost vs return calculation is made. As such, though the market may appear to be smaller, it is easier to engage with.

On the second front, understanding the bilingual requirements leads readily to multilingual capabilities where software products are developed. This is far more marketable. The issue here is the understanding and expectation in the market for what multilingual software is and how it should functionally perform. Though a compelling proposition can be made, in the world of software, having referenceable case studies is an essential requirement. In our case we're currently struggling to market our solutions outside of Wales due to a lack of interest and commitment within Wales (see points above).

To answer the question - yes, there is 'potential' for growth using SALT. However, our experience has shown that this still remains as potential and to use language solutions as part of a growth strategy can very much be a red herring and stifle growth due to a very weak market demand."

Richard Sheppard was also asked whether there was indirect potential for growth for SMEs in Wales, i.e. companies who used SALT components in other products"

"Yes, but I think this is very much the same answer as above since market demand is everything. Using SALT components in other products may be a different way of building the technology, but still requires the market demand to be there.

On a specific level, there are contrasting examples where, for instance, Cysgliad[15] is clearer marketable since it is targeted at the individual consumer who makes a purchasing decision around their unilateral language requirements. Whereas our efforts to build language technology into Enterprise solutions (CRM, corporate software, CMS, EDRMS, etc) has hardly any market interest. I suspect this is more to do with the type of

[15] Cysgliad is Bangor University's software compendium containing Welsh language spelling and grammar checking tools, thesaurus and a suite of electronic dictionaries and translation aids.

purchaser rather than whether it is direct language technology or embedded. Though, it could be argued that when language technology is embedded in other products that the purchasing decision has a broader range of criteria (i.e. does the package do x, y, z & also the language requirement) and it is then that we see the comparative importance of the language as a requirement that influences purchasing decisions."

Developing the market for SALT solutions

Asked what messages he would wish to convey to the Welsh Assembly Government through the SALT Cymru report, Richard Sheppard responded:

"Technology is great, but there needs to be a market. Until recently the argument was that the solutions weren't there. However, now that they are, unless the market makes use of them, the investment will have been wasted and there'll be no self-funded growth in these technologies."

In terms of his own company he replied:

"In our specific case, we think that WAG themselves could do more to embrace their needs. Without entering into specifics, we have seen a number of cases where bilingual requirements were stated in tender notices and then completely or largely disregarded as evaluation criteria. Also, there is often a poor understanding of what the requirements actually are and grossly inferior solutions are procured simply because the supplier had some 'Welsh' in the user interface.

This was similarly true of the whole sector: "we need the market to want language technology before it is worthwhile investing in it."

Skills and resource issues for SALT

Asked whether he would you welcome/need/use graduates with better SALT qualifications he replied:

"Of course. However, I would say our requirement is less specific than 'SALT'. Just 'T' (i.e. we need capable software engineers!) would suffice and we could teach the rest. This probably differs from CB [Canolfan Bedwyr] and other language technologists since we focus on functional support rather than linguistic.

Anyway, the net result/requirement is the same. Universities are not producing sufficient quantities and quality of employable output and this is a significant problem for us. It isn't so much about additional funding or strategic initiatives, just for universities to update their curriculum and standards."

Questioned on training needs for his existing workforce, in the form of workshops, mentoring activities, Richard Sheppard highlighted the fact that there seemed to be a great deal of training currently on offer, but that its quality was poor for their need, and that it did not produce the desired outcome for their company:

"There is an abundance of general and poor quality training provision. It seems to be easy to fund but poor in producing tangible outputs.

If there was high quality and specific training, then maybe so. However, I'm still not sure since in our sector, we are creators and core producers of technology. Commercial survival requires niche expertise - at which point, training is not feasible."

However, the idea of developing a freely available basic toolbox of resources for less-resourced languages such as Welsh up on the web for SMEs in Wales to use, and it was "always good to have resources available." The sustainability of 'free' resources was discussed, since there is always an economic cost to providing such resources. Draig themselves have provided some free resources on the web, including the To Bach utility (enabling UTF 8 compliant Welsh accented characters, including the problematic ŵ an ŷ to be correctly encoded in an electronic environment).

They would be happy to provide further such resources, given appropriate remuneration for their input:

"We have other ideas if someone wants to fund them, likewise we'd be happy to offer some SharePoint stuff for free. However, without a commercial benefit, not only would we be unable to offer the bulk of our SharePoint development (that has cost £100k+ to date) but we wouldn't also have the necessary revenue to continue further development."

Issues of quality control and maintenance were raised, especially the perception that 'free' often mean inferior in quality, and are often not maintained at project end:

"Free is hard to attach value to, and I would be concerned that quality and value of these resources is maintained, otherwise people wouldn't use them."

Creation of basic resources in the form of a suite of language tools that SMEs could customize and integrate into their own commercial products was a better idea. Many language utilities such as lemmatizers (allowing web searches for any form of a word, important in heavily inflected languages such as Welsh), named entities (such as names of people and places needed for data collection for speech recognition in e.g. call centres) and other 'building blocks' for developing bilingual and multilingual websites, databases etc. would be very useful to SMEs such as Draig, cutting not only the cost of development but also the time taken to develop an application from scratch:

"Very often we have to contend with tight deadlines in the delivery of our products, and pride ourselves in our 100% success rate in delivering within time and within budget. However, it is at present impossible to incorporate some language features into bilingual products as the basic modules are not yet available for Welsh. If the Welsh Assembly Government were willing to fund these and make them available to SMEs for further development in their own products, this would make a significant contribution to the software sector in Wales. SMEs would then be in a better position to deliver bilingual IT solutions to organisations that have a statutory obligation to maintain bilingual services in Wales."

Further needs

In conclusion, Richard Sheppard was asked what further needs Draig had identified for SALT and the IT sector in Wales. He replied as follows:

“On this thread the thing that I think is missing and I would love to see is a language resource portal that could consolidate all of our efforts and collaborations. There should be a public facing aspect, to inform people of what is available, what it all means, etc. There should also be an industry facing where we can collaborate, share knowledge, plan joint projects, discuss issues, etc. For instance, we have a wealth of knowledge around the Welsh Locale and we'd be happy to share this in a structure manner, etc.

There's us, CB [Canolfan Bedwyr], Meddal[16], Agored[17], BylG [the Welsh Language Board] and loads of others working in this area and loads of resources from your software and projects, our stuff, software standards and loads of other stuff and we all have disjointed websites and messages.... I've a bunch more thoughts around this, and though they're potentially exciting, they need a commitment and funding to create such a portal.”

He agreed that the establishment of a SALT Cymru special interest group would be a suitable vehicle to move forward with this agenda, and looked forward to be able to participate in such a group.

[16] Meddal is a voluntary group engadged in quality-controlled localization of software into Welsh, see http://www.meddal.com/english.htm

[17] Agored was an Objective 1/Welsh Language Board funded project at Aberystwyth University to develop a bilingual (Welsh/English) version of the free office tools OpenOffice.org 2.0, see http://agored.com/index_html_en

Appendix C5. Interview with Delyth Prys, Team Leader, Language Technologies Unit, Canolfan Bedwyr, Bangor University, March 28, 2008

Delyth Prys was asked to give an account of SALT related activities undertaken in the LTU

Background
SALT related activities began at BU in 1993 with the establishment of the Centre for the Standardization of Welsh Terminology at the School of Education. The work of the Centre introduced BSI and ISO standards for terminology and other language resources to Wales. In 2001 the Centre moved to Canolfan Bedwyr at Bangor University, and evolved into the Language Technologies Unit (LTU) when the team expanded to include software developers and speech technologists. This mirrors the development of the ISO Technical Committee (TC 37) responsible for Terminology Standards, which has expanded to develop standards for other language resources given their increasing significance in digital and multimodal communications and the advent of the internet. The Unit is entirely self-funding, and won £1 million in research grants and commissions during the period 2001-2007. The LTU's four main areas of activities are:

1. Terminology standardization
2. Place-name research
3. Language tools for software
4. Speech technology

These map well onto the 8 SALT categories identified in 5.4 of this report, with only the second of the 8 categories (Written language input in the form of optical character recognition and handwriting recognition) so far not covered in the LTU's activities.

Terminology Standardization
The LTU's team leader, Delyth Prys, and Dewi Bryn Jones, are members of the BSI (British Standards Institute)'s Technical Committee on Terminology and Other Language Resources. This is the committee which represents the UK on the corresponding ISO (International Standards Organisation)'s Technical Committee. The BSI's interest in terminology standardization stems from its importance to the development of business information solutions for British organizations of all sizes and sectors. BSI British Standards works with manufacturing and service industries, businesses, governments and consumers to facilitate the production of British, European and international standards.

The LTU has developed standards and guidelines for Welsh terminology work, and in 1998 and 2007 was commissioned by the Welsh Language Board to write Guidelines for the Standardization of Terminology[18]. To date the Unit has produced over 20 dictionaries of standardized terms for various organisations in Wales and beyond. It has standardized terminology for, amongst others: ACCAC, Local Health Boards, the Environment Agency Wales, the Electoral Commission, Microsoft and the Welsh Language Board[19].

Although this focus is on technical terminology (defined as a concept-based, prescriptive science) rather than on general lexicography (understood as a word-based, descriptive

[18] See http://www.bwrdd-yr-iaith.org.uk/cynnwys.php?pID=109&langID=2&nID=2823
[19] See http://www.bangor.ac.uk/ar/cb/termau.php.en?catid=&subid=3279 for a bibliography of its published dictionaries.

discipline) there is increasing cross-over between terminology and general lexicography work, especially in the use of software tools to manage and develop both terminology and lexicography projects. The Unit has therefore undertaken some general lexicography projects, including an on-line Welsh-Irish phrase-book and dictionary, and currently a digital interface to enable a dispersed team of editors to produce an updated electronic version of the Welsh Academy Dictionary together, commissioned by the Welsh Language Board.

The Unit has developed tools for the manipulation of terminological data, enabling it, amongst other things to publish dictionaries from its databases simultaneously in paper and electronic formats. The electronic formats include CD ROMs, interactive searchable web dictionaries, and downloads for mobile phones. It is increasingly aware that standardized terminologies are the basis for many advanced applications for knowledge management. Concept-based terminology work is closely related to semantics, controlled vocabularies and IPSV. Such controlled lists needed to populate subject and other metadata in human and machine readable formats, and for the development of the semantic web itself. Although it understands terminology standardization to be a valid activity in a monolingual environment, it foresees much of its future activity being undertaken in an increasingly bilingual and multilingual framework, with Welsh being mainstreamed in an international multilingual, multimodal and multimedia context.

Place-name research
The LTU holds two major databases of Welsh place-names, one historical[20] and one contemporary[21]. It also engages in place-name research for others, such as articles for the BBC 'What's In a Name' project[22] in 2007.

Place-name data is an important component in many new technology applications, and is crucial for many business services such as call-centres and postal deliveries. Linking information in the databases to on-line maps has been a much welcomed recent addition to the place-names database, and further work on visualization of this data is anticipated.

There has been increased appreciation of the importance of place-name data as named entities in speech recognition work, and much of this data has been reused in the speech technology development (see below). The LTU's Llefaru Lleoedd project aims to produce a commercial list of Welsh place-names for use by industry in various speech applications.

Language Tools for Software
The LTU has produced Welsh language office tools for Microsoft (spellchecker and hyphenator) and OpenOffice (spellchecker). Its own commercial suite of office tools, Cysgliad (spelling and grammar checker, thesaurus, word-by-word translator, compendium of general language and technical terminology dictionaries) is available both in home and business editions. The business edition is much used by public and private organisations in Wales to facilitate operating in a bilingual environment.

[20] Archif Melville Richards Place-name Database, see http://www.e-gymraeg.co.uk/enwaulleoedd/amr/
[21] Enwau Cymru database, see http://www.e-gymraeg.co.uk/enwaucymru/
[22] See http://www.bbc.co.uk/wales/whatsinaname/

Cysgliad is powered by Cyslib, a library of Welsh linguistic utilities, resources and software components developed at the LTU. These include a Part of Speech Tagger and a Lemmatizer which are key components of many applications, including those language analysis and intelligent web searches. A number of these are now being licensed individually to industry (many of them SMEs in Wales) for inclusion in their own products, on a B2B model.

Software applications developed by the LTU for language learning have also been able to use the Cyslib suite. These include the dictionary, spell checker and mutation checker on the BBC's Learn Welsh website[23], and the Ieithgi[24] and some of the RAW games[25], also commissioned by the BBC.

There is considerable overlap on content research and software development. This has enabled the team's software specialists to develop applications where form and content are closely integrated. Facilitating the future reuse of both code and content is an important part of the LTU's philosophy, as is expanding from servicing the bilingual requirements of Wales to embrace a broader multilingual global perspective.

Speech Technology

This was initiated in the LTU in the WISPR project. [26] This was a joint research Interreg IIIA–funded project with the Dublin universities (Trinity College Dublin, University College Dublin and Dublin City University). Additional funding for the Welsh language work was provided by the Welsh Language Board.

This project provided a workable model of joint cooperation between university departments which successfully delivered basic speech resources for both Welsh and Irish. It also pioneered a successful model for engaging with industry where basic resources were made freely available under a BSD style licence, thus enabling SMEs to incorporate and/or further develop speech components in their commercial products. It also enabled the LTU to further develop the resources itself, and to market a commercial voice deployed to voice-enable many websites.

In April 2008 the LTU will begin a new Welsh Language Board funded project to develop basic speech recognition resources for Welsh. The output from this project will be made freely available for further development by industry following the successful model used in the WISPR project.

KTPs and Knowledge Transfer

In 1998 the LTU (then known as the Centre for the Standardization of Welsh Terminology) was awarded its first knowledge transfer project with Cymen, a translation company based in Caernarfon, Gwynedd. This led to the introduction of language technology tools, including Translation Memory, into the company, enabling it to stay ahead as one of the leading translation companies in Wales. Cymen, with their new language technology expertise, went on to win the Microsoft contract for localizing Microsoft Windows and Microsoft Office suite into Welsh.

[23] http://www.bbc.co.uk/wales/learnwelsh/level_test/
[24] http://www.bbc.co.uk/cymru/ieithgi/
[25] http://www.bbc.co.uk/cymru/raw/gemau/
[26] http://www.bangor.ac.uk/ar/cb/wispr.php.cy

Since this time the LTU has been active in its support of Welsh SMEs. It has participated in another KTP with Y Lolfa, a Welsh publishing house based in Talybont, Ceredigion. It provides training for the Translation sector, through both academic and industry channels. It provides advice and support for SMEs seeking to use its utilities in their products.

In 2008 it was awarded a KTP with Testun, a company specialising in translation, subtitling and teletext services based in Cardiff. This project will concentrate on developing speech recognition and machine translation modules for the company's own internal use and for sale to other potential users.

Appendix D. The SALT Cymru survey – results and analysis

Introduction

The SALT Cymru survey was devised as a means of gathering data about the state of Speech and Language Technologies (SALT) development in Wales. Although primarily concerned with developers or potential developers of SALT in Wales, the survey also collected information regarding the use SALT amongst users of the technology.

Method

Design

As a specialized survey, the SALT Cymru Survey was not targeted at the general public but rather at those in industry and public organisations. So as to maximize the number of respondents, the survey avoided excessive length and was kept as straightforward as possible. Questions rendered irrelevant to the respondent by their earlier answers were not displayed, and guides and definitions accompanied questions so as to clarify and explain the terminology used. Survey respondents could save their responses mid-session and return to them at a later date, and the respondents could participate anonymously if they so desired.

Delivery

The survey was created using Limesurvey, an open source online survey software program which was installed and hosted on the Language Technologies Unit's own servers for the duration of the survey. Limesurvey was chosen as it offered a wide range of professional grade survey tools as well as the capability for both online and offline (printed format) delivery. In addition, Limesurvey supported multilingual content, and its open source nature had the added advantage of minimizing cost whilst also allowing the survey interface to be translated into Welsh (this translation will be made available for the benefit of all at the end of the project). Limesurvey also allowed the SALT Cymru Survey to be customized with the branding of the SALT Cymru project and that of the Welsh Assembly Government, enhancing the survey's professional look and feel.

Marketing

The survey was marketed in several different ways: a large number of invitations were sent out to specially targeted key players, organizations and companies. These recipients were targeted on the basis of their relationship to SALT and SALT's potential relevance to their work, as ascertained through research carried out by the SALT Cymru project's researchers. A form on the front page of the SALT Cymru website also invited visitors to the website to register their interest, and those who did were sent an invitation to complete the survey. The website itself was marketed as part of the project's general marketing, with the web address appearing in articles concerning the project which appeared in both general publications such as The Daily Post and specialized publications such as Advances Wales. In addition, promotional literature such as leaflets and posters prominently displayed the website address, and both the website and the survey was marketed at exhibitions and conferences attended by the SALT Cymru team.

Results

The results of the SALT Survey appear below. Rather than follow the question order found in the survey, the results are displayed in a manner intended to be more

meaningful to the reader. The full list of questions and their original order of presentation is given on the following two pages. Note that as some questions were only triggered by specific responses from the respondent, not all questions would appear during the course of responding to the survey.

Survey Questions

0001: A1. Are you completing this survey as an individual, or on behalf of an organization?

if organization:

0005: A3. How many people are employed in your organization?

0002: A4. Do you have a website?

if so:

0006: B1. Is your website available in more than one language?

if so:

0007: B2. In what language or languages do you provide your website?

0008: C1. Do you develop SALT?

if so group F (0012-0017) must be answered

0009: C2. Do you use SALT?

if so:

0010: D1. How often, if at all, do you use these parts of speech and language technology (SALT)?

if not, but would like to:

0011: E1. In what context or contexts would you like to use SALT?

SALT DEVELOPERS QUESTIONS

0012: F1. What SALT products or techniques do you currently develop, or have you developed in the past?

0013: F2. What SALT do you intend to develop in the future?

0014: F3. What other technologies do you develop?

0015: F4. What is your main source of funding?

0016: F5. At what markets do you target the SALT that you develop?

0017: F6. How important are the following aspects of SALT to you?

PROSPECTIVE SALT DEVELOPERS QUESTIONS

0018: G1. What SALT are you interested in developing in the future?

0019: G2. At what markets would you target the SALT that you would develop?

0020: G3. What other technologies do you currently develop?

0021: G4. What is your main source of funding?

QUESTIONS FOR NON-INTERESTED NON SALT-USERS

0022: H1. Reasons (for no interest in SALT):

RESPONDENT DETAILS

0029: Are you happy to be contacted with further information about the SALT Cymru project?

0023: Your first name

0024: Your surname

0025: Your contact telephone number

0026: Your email address

0027: Your postal address

0028: Your post code

*Note that these were internal question numbers only, and that some numbers were not utilized by the final survey. There were no questions numbered ‘3’ and ‘4’ in the final survey, for example.

1. (0008: C1.) Do you develop SALT?

i) Results for: All Respondents (48 of 48 respondents)

From the many who registered their interest and received an invitation to complete the survey, a total of 48[27] respondents completed the SALT Cymru Survey. This is considered a satisfactory figure in light of the specialized nature of the subject and the project's scope and timeline.

As was expected from their willingness to complete the survey, most of the survey respondents professed an interest in SALT, either as developers, prospective developers or users, with only three responding that SALT was of no interest to them[28].

The respondents were grouped by their relationship to SALT, as follows:

> **40% SALT Developers** – these are survey respondents who have indicated that they are developers of speech and language Technology (SALT)
>
> **31% Prospective SALT Developers** – these are survey respondents who have indicated that, although they do not currently develop SALT, they may be interested in doing so in the future.
>
> **29% Non-developers of SALT** – these are survey respondents that are not SALT developers and do not intend to develop SALT in the future.

0008: C1. All Respondents - Do you develop SALT?

[27] A further 20 respondents failed to complete the survey in its entirety, including some who seem to have mistaken the SALT acronym for *Speech and Language Therapy*.
[28] Again, there is an indicaton that these were respondents that had confused Speech and Language Technology with Speech and Language *Therapy*.

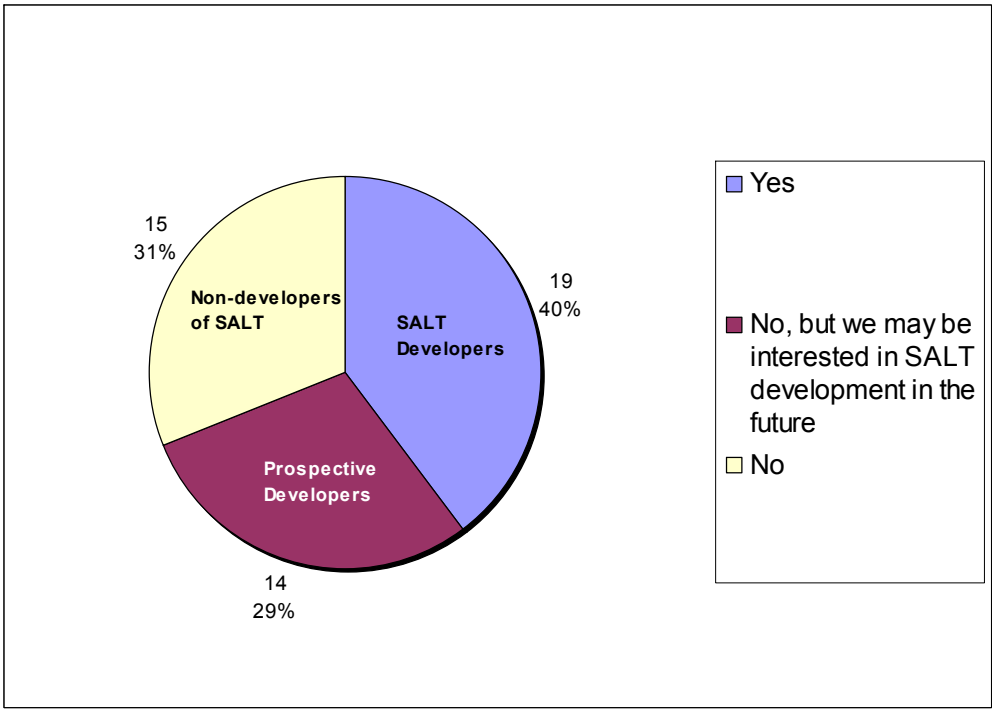


Table: 0008: C1. All Respondents - Do you develop SALT?

Yes	19
No, but we may be interested in SALT development in the future	14
No	15
	48

The above results demonstrate that respondents from each of the three categories are well represented in the SALT Cymru survey. The correspondence between **Prospective SALT Developers** and **Non-developers of SALT** is close to equal, whilst the largest proportion of respondents consisted of those identifying themselves as **SALT Developers**, who were the survey's major target.

Considering the specialized nature of SALT, the number of respondents indicating an interest in becoming SALT developers (i.e. Prospective SALT Developers) is high, indicating that SALT is perceived as representing future opportunities to those not currently involved in the field.

The number of respondents responding to the survey despite having no interest in becoming SALT developers (Non-developers of SALT) suggests that developments in Speech and Language Technologies are of interest to users in general, and not only to those involved in developing SALT themselves.

1. 0001: A1. Are you completing this survey as an individual, or on behalf of an organization?

i) Results for: All Respondents (48 of 48 respondents)

Of the 48 survey respondents, 19 (40%) were *Companies or Commercial Organizations*, 17 (35%) were *Private Individuals* and 8 (17%) were from *Higher Education Establishments*. One respondent each (6%) replied for the following categories: *Health or Care Establishment*, *Not-for-profit Organization or Charity* and *Other*.

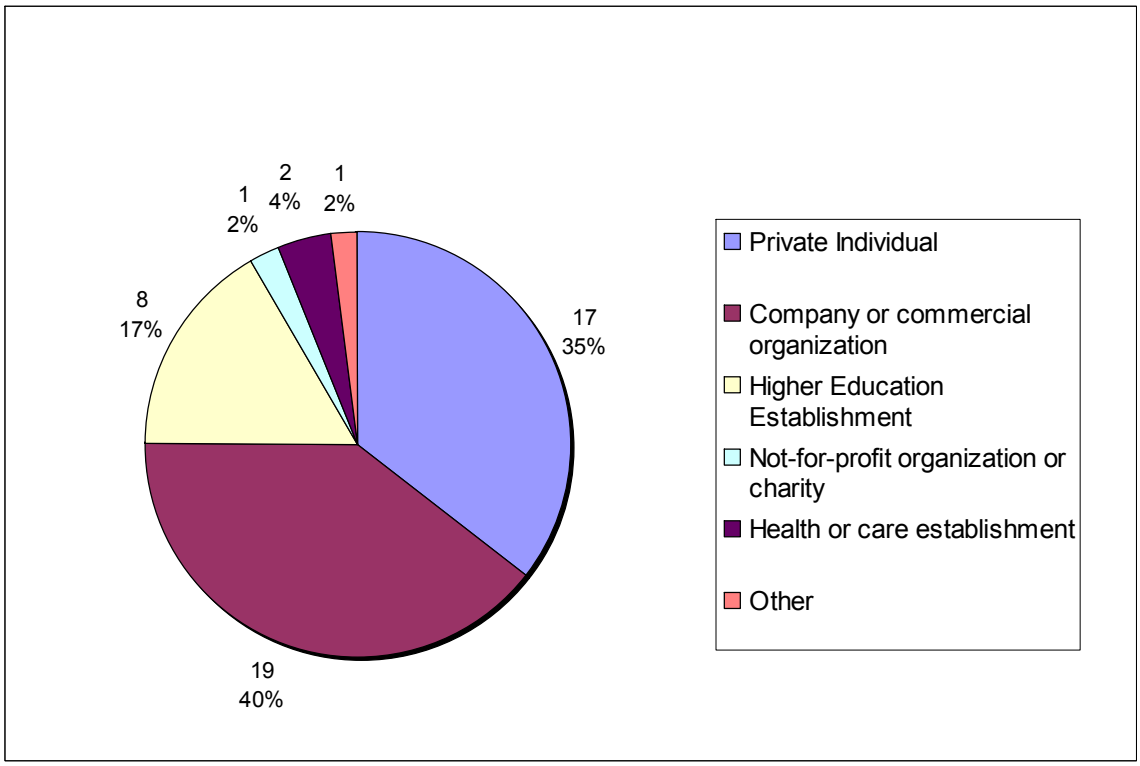


Table: 0001(i): A1. All Respondents - Are you completing this survey as an individual, or on behalf of an organization?

Private Individual	17
Company or commercial organization	19
Higher Education Establishment	8
Not-for-profit organization or charity	1
Health or care establishment	2
Other	1
	48

While responses from Private Individuals represented a large proportion of the respondents, it should be remembered that these responses represent single

individuals, whilst those received from organizations and establishment are often representative of teams comprising more than one individual.

I should also be noted that, although the *Higher Education Establishment* was placed in third with 8 responses (17%), this response in fact represents a significant proportion of the HEIs in Wales. Responses from the *Not-for-profit Organization or Charity* grouping was disappointing considering the potential of SALT to address issues concerning the access of disabled and elderly users to technology.

1. (0001: A1.) Are you completing this survey as an individual, or on behalf of an organization? (continued)

ii) Results for: SALT Developers (19 of 48 respondents)
Of the 19 survey respondents who indicated that they were **SALT Developers**., 7 (37%) were *Companies or Commercial Organizations*, 6 (24%) were from *Higher Education Establishments*, and 3 (16%) were *Private Individuals*. One respondent (6%) each replied for the following categories: *Health or Care Establishment*, *Not-for-profit Organization or Charity* and *Other*.

0001: A1. SALT Developers - Are you completing this survey as an individual, or on behalf of an organization?

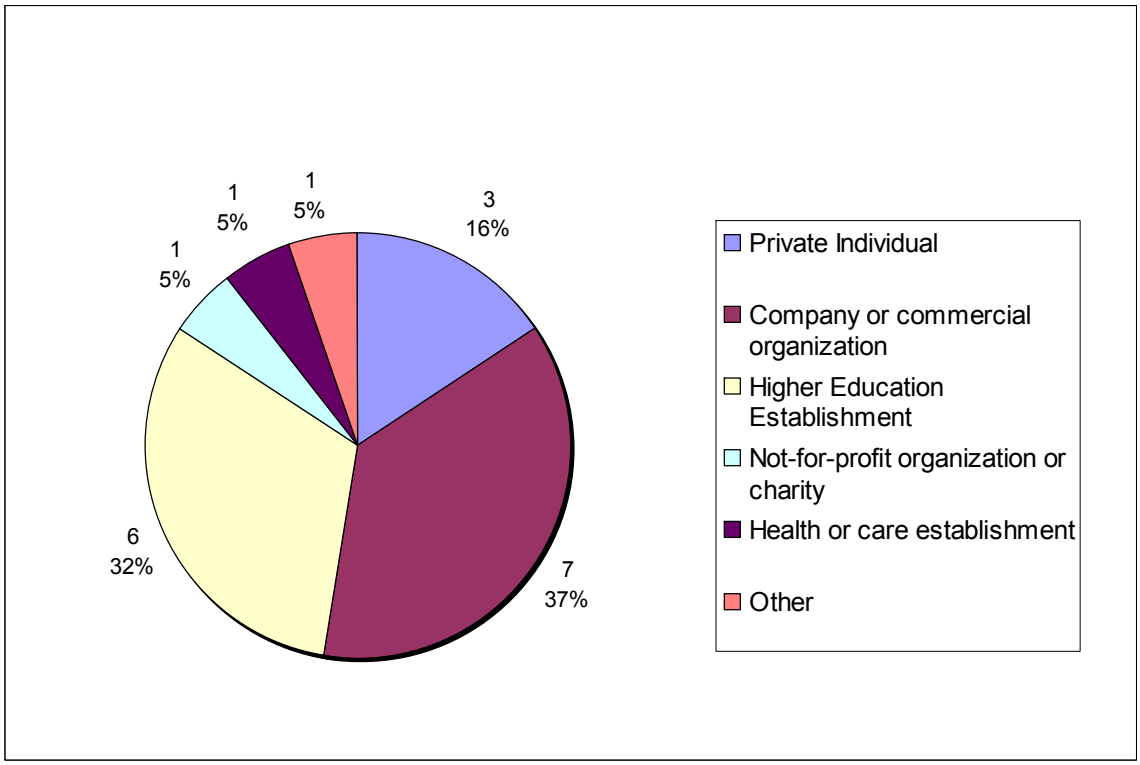


Table: 0001(ii): A1. SALT Developers - Are you completing this survey as an individual, or on behalf of an organization?

Private Individual	3
Company or commercial organization	7
Higher Education Establishment	6
Not-for-profit organization or charity	1
Health or care establishment	1
Other	1
	19

The above chart shows that the majority of survey responses from **SALT Developers** came from those representing *Companies or Commercial Organizations*.

The second largest category was that of *Higher Education Establishments*, representing academic research carried out at Welsh universities. Whilst only accounting for 18% of **All Respondents**, *Higher Education Establishments* accounted for 32% of **SALT Developers**, demonstrating their positions as important centres of research and development.

Third in terms of proportion was the category of *Private Individuals*, perhaps representing self-employed individuals who did not perceive themselves as a 'company' or 'commercial organization', individuals looking to start commercial enterprises in the field of SALT, or competent non-professionals contributing to open source development.

The *Health or Care Establishment* was represented by a single respondent, and similarly one respondent indicated that they belonged to the *Not-for-profit organization or charity* category. Another respondent was marked as '*Other'*.

1. (0001: A1.) Are you completing this survey as an individual, or on behalf of an organization? (continued)

iii) Results for: Prospective SALT Developers (14 of 48 respondents)

Of the 14 survey respondents who indicated that they were **Prospective SALT Developers**, 7 (50%) were *Private Individuals*, 4 (29%) were from *Companies or Commercial Organizations*, and 2 (14%) were *Higher Education Establishments*. One respondent (7%) replied from *Health or Care Establishment*, *Not-for-profit Organization or Charity* and *Other*.

0001: A1 Prospective SALT Developers - Are you completing this survey as an individual, or on behalf of an organization?

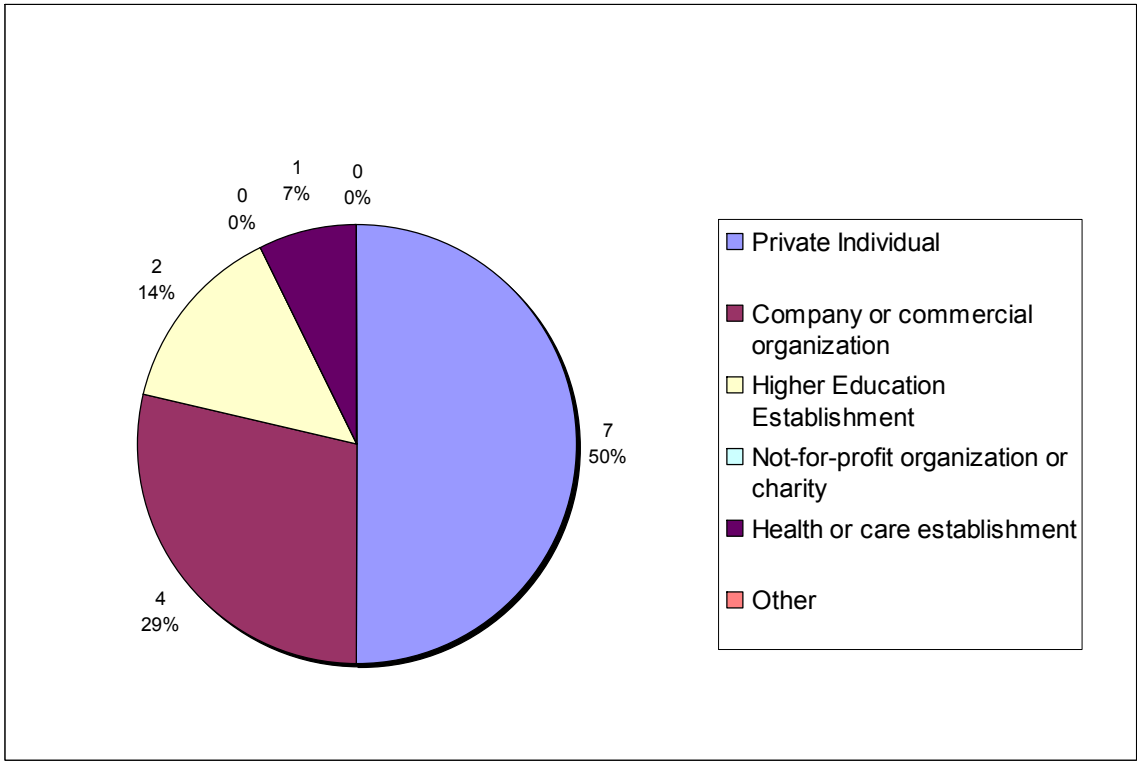


Table: 0001(ii): Prospective SALT Developers - Are you completing this survey as an individual, or on behalf of an organization?

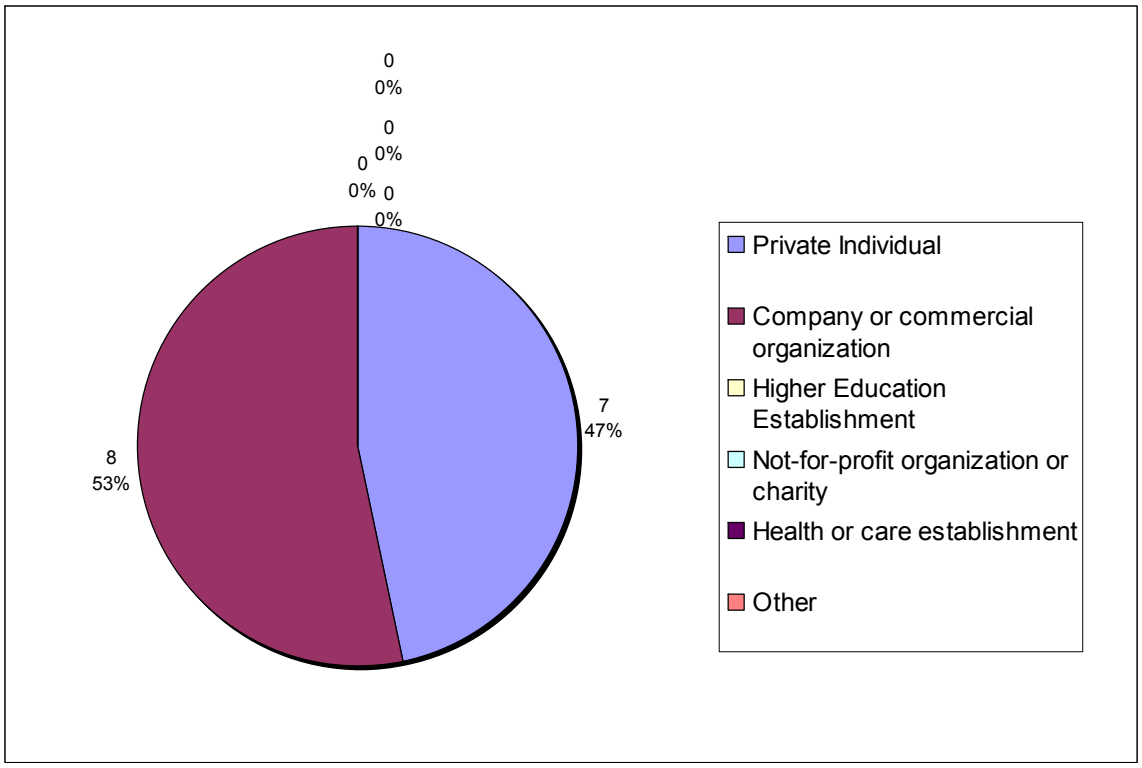
Private Individual	7
Company or commercial organization	4
Higher Education Establishment	2
Not-for-profit organization or charity	0
Health or care establishment	1
Other	0
	14

Those indicating that they would be interested in becoming SALT Developers were primarily **Private Individuals**. This seems to indicate that many of the Private Individuals responding to the survey considered themselves capable of developing SALT, either as individuals, or by joining larger SALT development teams. A single respondent from a Health or Care establishment again signified an interest in SALT development from that sector.

1. (0001: A1.) Are you completing this survey as an individual, or on behalf of an organization? (continued)

iii) Results for: Non-developers of SALT (15 of 48 respondents)

Of the 15 Non-developers who responded, 8 (53%) were *Companies or Commercial Organizations*, 7 (47%) were *Private Individuals*.



0001: A1. Non-Developers of SALT - Are you completing this survey as an individual, or on behalf of an organization?

Private Individual	7
Company or commercial organization	8
Higher Education Establishment	0
Not-for-profit organization or charity	0
Health or care establishment	0
Other	0
	15

Above, we see that respondents not involved in SALT development and not interested in becoming SALT developers in the future were either *Companies or Commercial Organizations* or *Private Individuals*. The absence of *Higher Education Establishments* here indicates an involvement or interest in SALT development from all those who

replied from the Higher Education sector. Other sectors such as *Health or Care Establishments* seem to be absent from this category due to not having responded to the survey in sufficient numbers.

2. (0005: A3.) How many people are employed in your organization?

Number of employees employed by respondents indicating that they were a Company or Commercial Organization (19 of a total of 48 respondents)

Of the 19 survey respondents who identified themselves as representing *Companies or Commercial Organizations*, 10 (52%) employed *2-9 employees*, 6 (32%) employed a single employee and 3 (16%) employed 50-249 employees. None of the respondents represented *Companies or Commercial Organizations* employing 10-49 or 250+ employees.

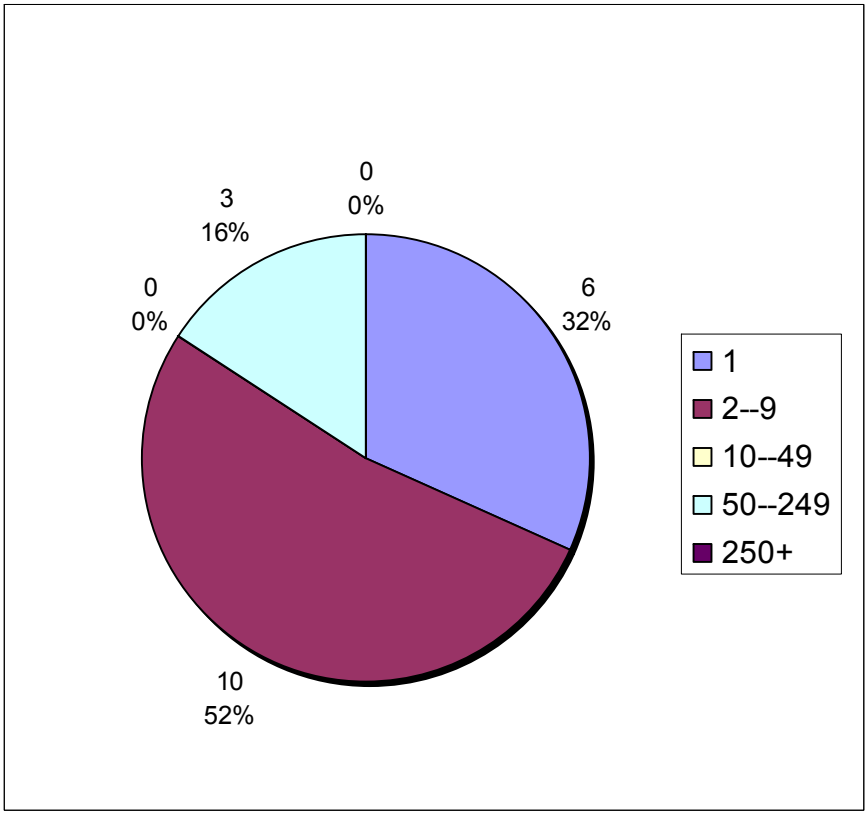


Table 0005: A3. (i) Companies or Commercial Organizations – How many people are employed in your organization?

1	6
2--9	10
10--49	0
50--249	3
250+	0
	19

Respondents identifying themselves as *Companies and Commercial Organizations* were asked to state the size of their organization in terms of the number of employees employed. The available options corresponded to the EU definition of SMEs according to headcount, which can be found at the following website:

http://ec.europa.eu/enterprise/enterprise_policy/sme_definition/index_en.htm)

For convenience, a summary is reprinted below:

Enterprise category	Headcount
medium-sized	< 250
small	< 50
micro	< 10

Note that the survey results above are slightly more detailed than the categories outlined by the EU, as companies or commercial organizations employing a single employee have been categorized separately. Below are the same results combined into categories that correspond exactly to those of the EU:

Number of employees employed by respondents indicating that they were a Company or Commercial Organization (19 of a total of 48 respondents)

Of the 19 survey respondents who identified themselves as representing *Companies or Commercial Organizations*, 16 (84%) employed 1-9 employees and 3 (16%) employed 50-249 employees.

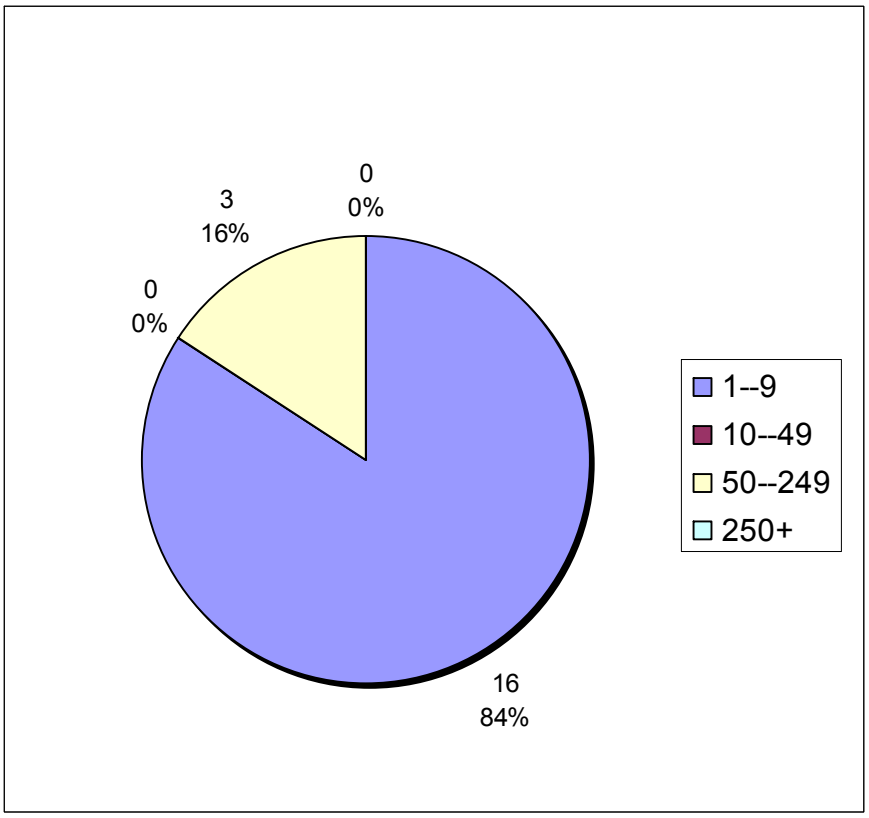


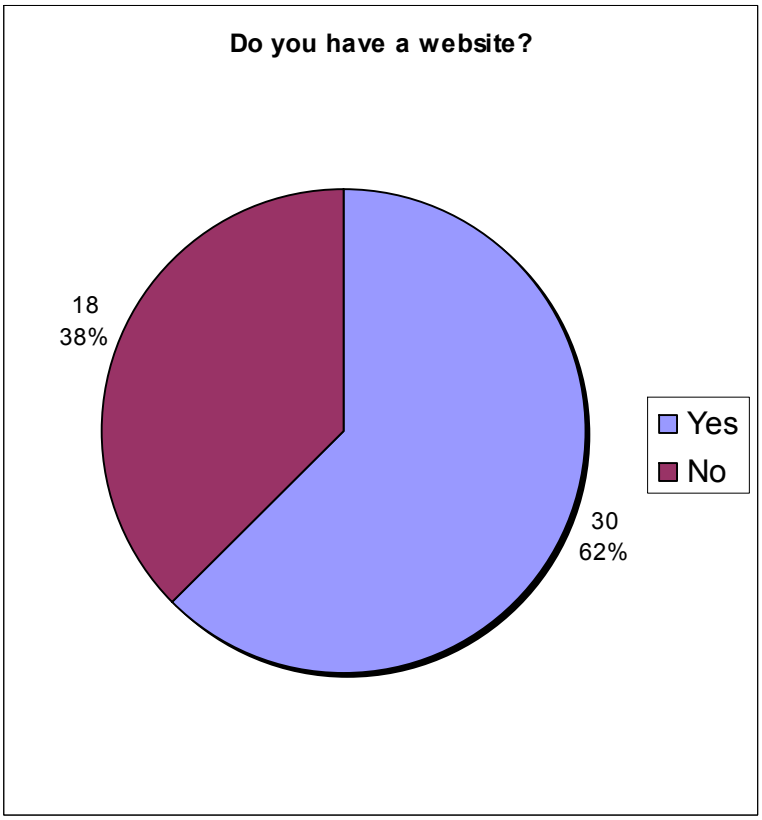
Table 0005: A3. (ii) Companies or Commercial Organizations – How many people are employed in your organization?

1-9	16
10-49	0
50-249	3
250+	0
	18

As can be seen above, the vast majority of *Companies or Commercial Organizations* responding to the survey are *micro SME* sized (84%), with only two companies belong to another category, that of the *medium-sized SME* (50-249 employees) (16%). However, over a third of the *micro SME* sized organizations employed a single employee only.

3. (0002: A4.) All Respondents - Do you have a website?

Results for: All Respondents (48 of 48 respondents)
Of the 48 survey respondents, 10 (62%) employed 2-9 employees, 30 (62%) reported that they possessed a website and 18 (38%) reported that they did not.



Yes	30
No	18
	48

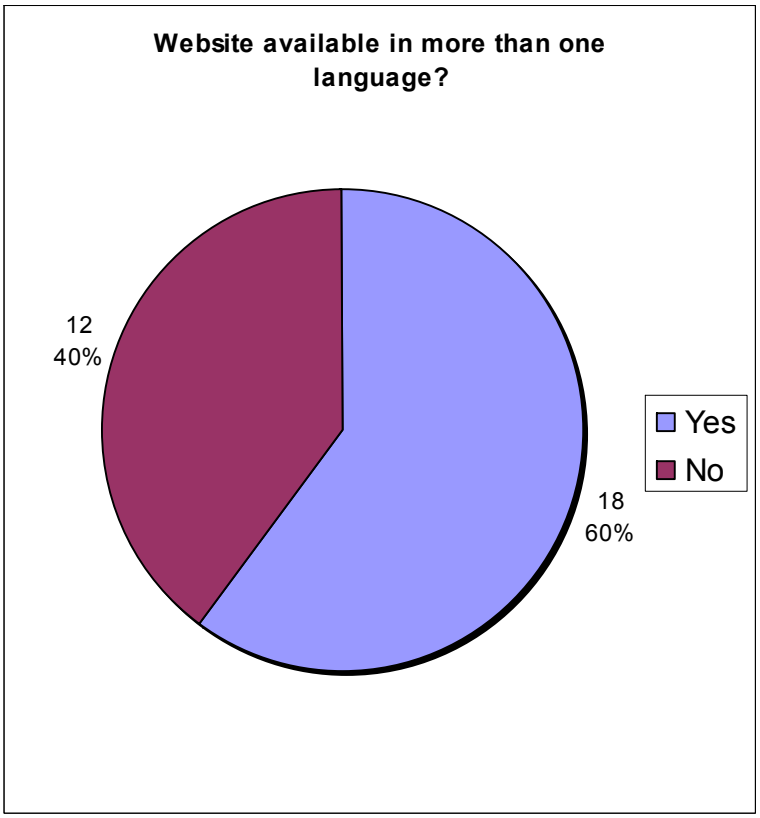
Data on whether respondents possessed websites was collected by the survey as the possession of websites served as an indicator of respondents' web presence in addition to their technical capability. Those with websites were asked to supply their website addresses, so that further data about those who had responded to the survey could be gathered if needed. In some cases, data to be found on a respondent's website allowed the validation or correction of information which appeared to be incomplete or erroneous.

The fact hat 62% of the survey respondents indicated that they possessed a website demonstrates the importance of the web to the survey respondents. Those who classified themselves as SALT Developers responded that 89% possessed websites,

demonstrating an increased emphasis on the web and a higher technological capability amongst developers.

4. (0006: B1.) Website owners - Is your website available in more than one language?

Results for: All Respondents possessing websites (30 of 48 respondents)
Of the 30 survey respondents who possessed websites, 18 (60%) possessed multilingual websites and 30 (62%) possessed monolingual websites.

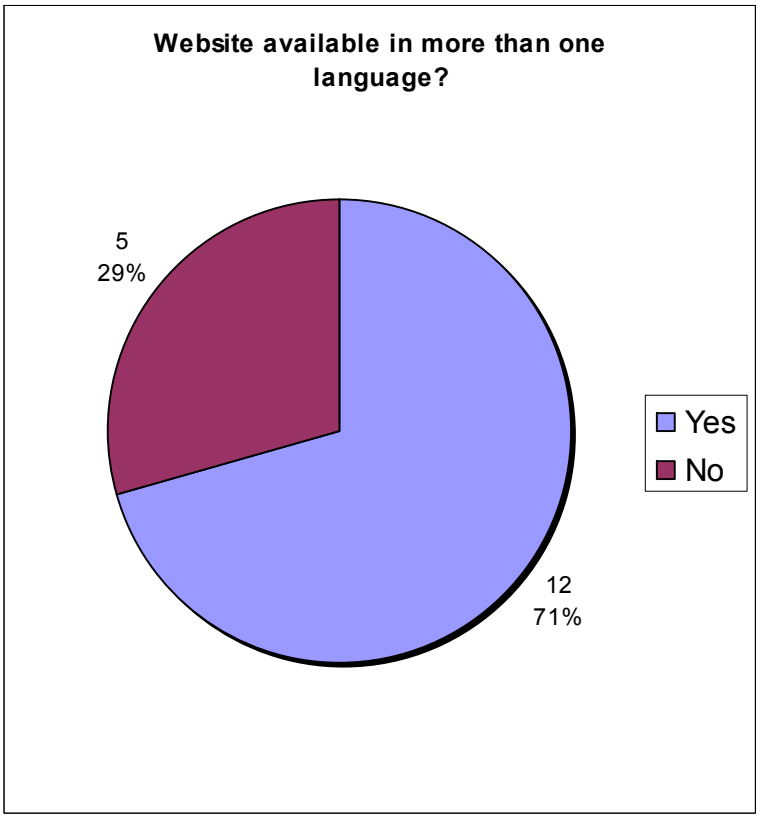


Yes	18
No	12
	30

The survey collected data on the proportion of respondents whose websites were multilingual. A multilingual website was seen as an indicator of how important multilingualism was to the respondent.

As can be seen from the above results, a large percentage (60%) of the websites belonging to the survey's respondents were multilingual, demonstrating the importance of multilingualism to a large proportion of respondents, and showing that the desire and ability to support multilingualism in a computing environment is widespread.

Of the various categories of respondents, a higher proportion of multilingual websites belonged to SALT Developers:

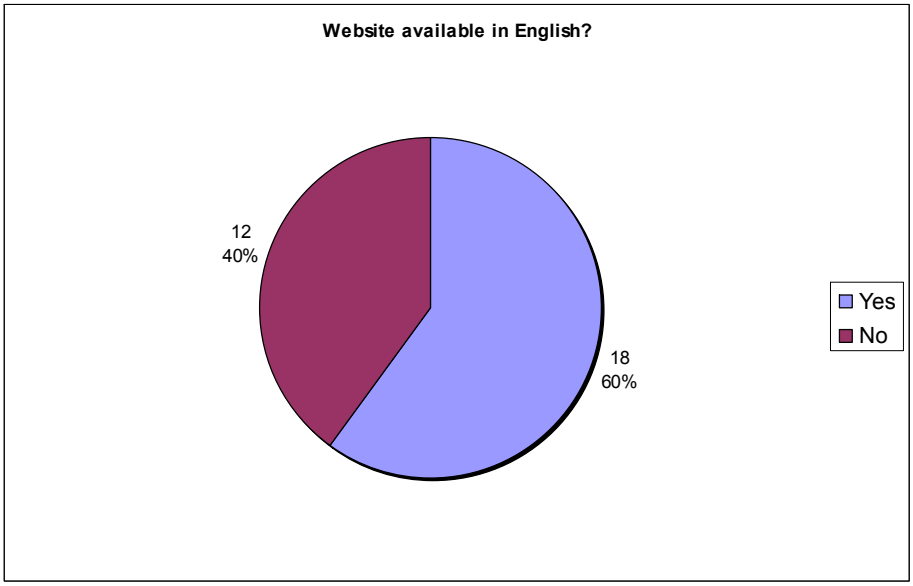


Yes	12
No	5
	17

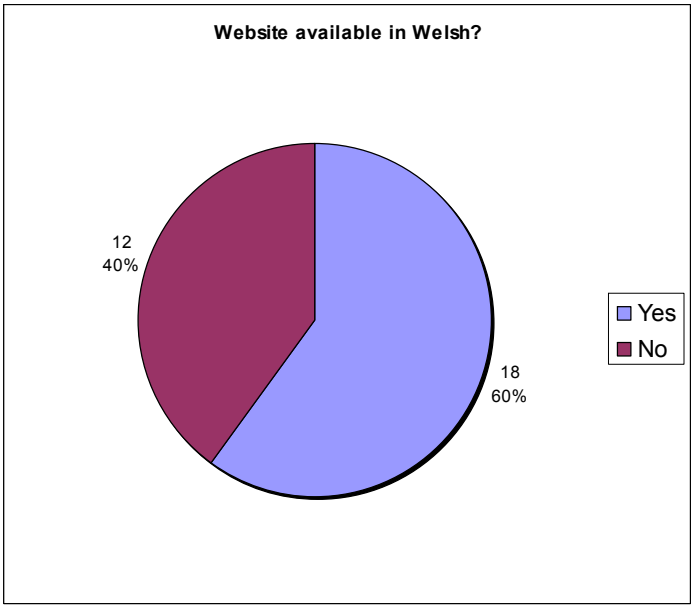
Websites owned by SALT Developers possessed a greater degree of multilingualism (71%), indicating that developers placed a greater emphasis on multilingualism than non-developers.

5. (0007: B2.) Website owners - In what language or languages do you provide your website?

Results for: All Respondents possessing websites (30 of 48 respondents)



Yes	18
No	12
	30



Yes	18

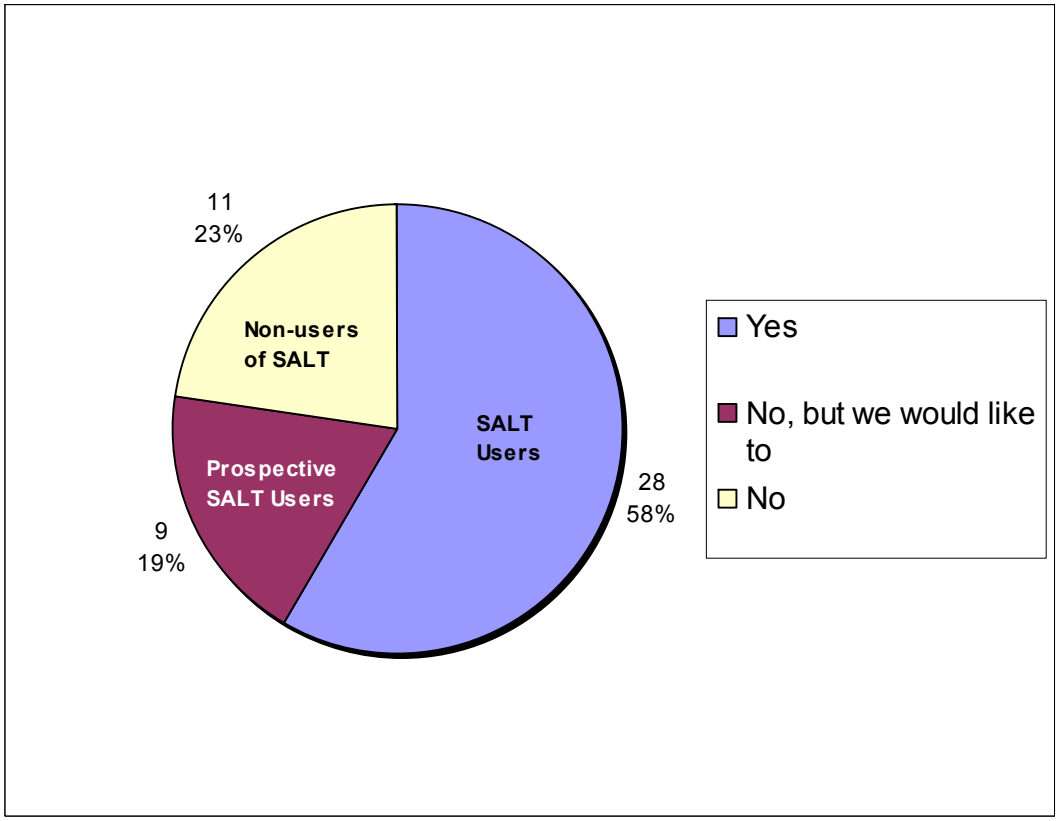
No	13
	30

Of the 30 survey respondents who possessed websites, 18 (60%) possessed websites which were available in English and 12 (40%) did not. Similarly, 18 (60%) possessed websites which were available in Welsh and 12 (40%) did not. Note that these 30 websites include bilingual English and Welsh websites as well as websites which are available in English or Welsh only.

All the multilingual sites reported were bilingual sites offering the choice of either English or Welsh, except for one trilingual site that also offered French as an option. Multilingualism amongst respondents to the SALT Survey therefore seems to currently stem from the bilingual situation in Wales rather than from an international or global multilingualism.

6.(i) (0009: C2.) All Respondents - Do you use SALT?

Results for: All Respondents (48 of 48 respondents)
Of the 48 survey respondents, 28 (58%) indicated that they used SALT, 9 (19%) indicated that did not use SALT (but would like to) and 11 (23%) responded that they did not use SALT (without indicating that they would like to).

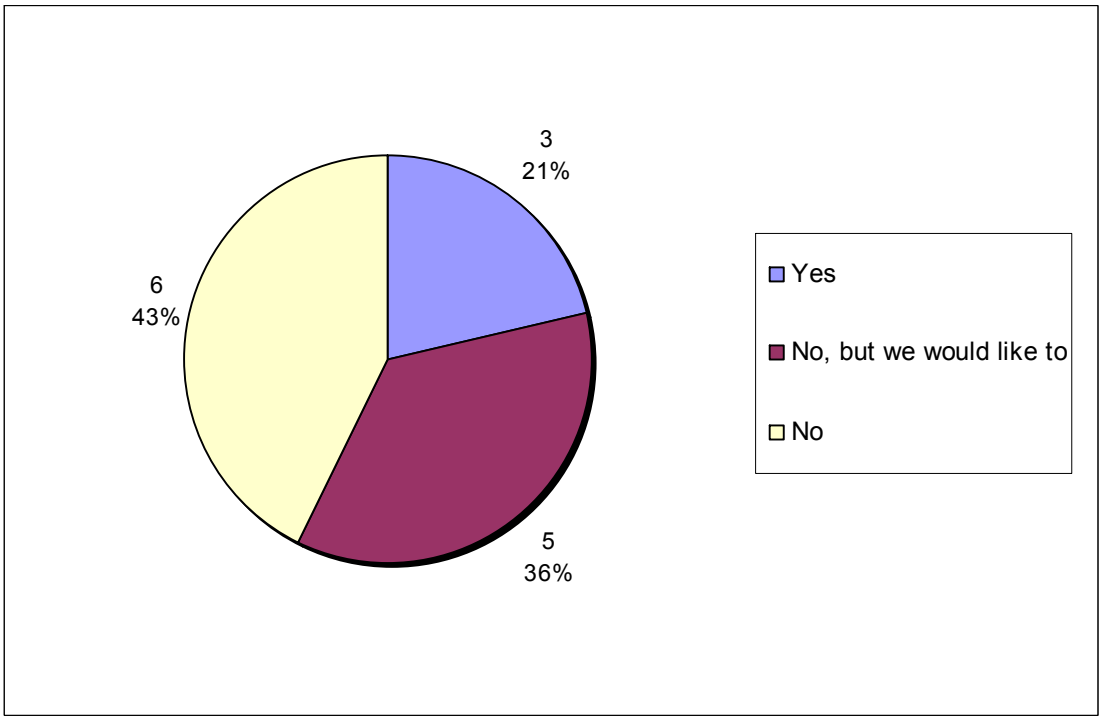


Yes	28
No, but we would like to	9
No	11
	48

Although the majority of respondents (58%) indicated that they use SALT, this figure is lower than expected given the survey's targeted recipients. Considering that such ubiquitous technologies such as spellcheckers and online dictionaries are included within the definition of SALT technologies, most if not all the respondents would have been expected to make some use of SALT technology. It would seem that this result reflects a general confusion regarding the definition of what constitutes 'Speech and Language Technology' amongst those not actively involved in its development.

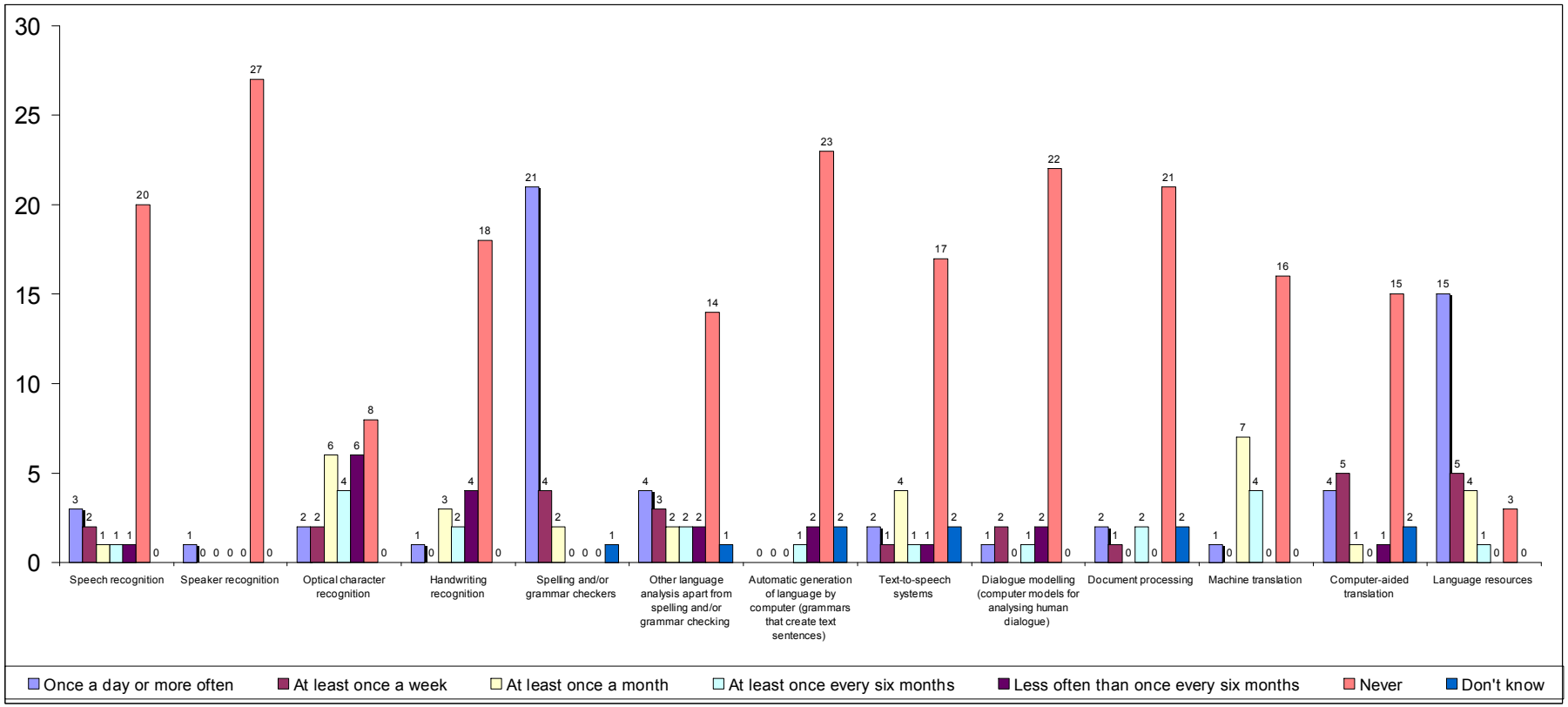
6.(ii) (0009: C2.) Prospective Developers - Do you use SALT?

Results for: Prospective Developers (14 of 48 respondents)
Of the 14 Prospective Developers, only 3 (21%) indicated that they used SALT, 5 (36%) indicated that did not use SALT (but would like to) and 6 (43%) responded that they did not use SALT (without indicating that they would like to in the future).



Whilst the vast majority of **SALT Developers** claimed to make use of SALT technology (84%), or wished to (11%), only (21%) of those interested in developing SALT in the future (i.e. Prospective Developers) indicated that they were users of SALT. In fact a surprising number (43%) of Prospective **Developers** chose not to indicate that they desired to use SALT in the future, whilst 36% indicated they were not SALT users but that they would like to make use of SALT in the future. This perhaps reflects a belief amongst those not immediately familiar with SALT development that SALT technologies refer only to speech and language software development kits, and not to technologies bought at retail and used by the ordinary user. It could also reflect Prospective Developers involved in web design who could envisage developing websites that include accessibility features catering for disabled users, but who would not use the features themselves as users.

7. (0010: D1.) Salt Users - How often, if at all, do you use these parts of speech and language technology (SALT)?



7. (0010: D1.) SALT Users - How often, if at all, do you use these parts of speech and language technology (SALT)?

Immediately apparent from the bar chart above is that many of the survey respondents indicated that they have never used SALT technologies from many of the more specialized categories above. For example, the only respondent to indicate any use of Speaker Recognition was a developer of Speaker Recognition software, which is not surprising considering Speaker Recognition is not yet in widespread use. However, in nearly all these cases there is a small but significant number of respondents who use these technologies on a daily basis, indicating that to a section of the survey respondents these are very important technologies.

Technologies such as spellcheckers and language resources which are widely used and recognized by general users are well represented in the above results. These are technologies that often come preinstalled on computers and are employed by users working on general everyday tasks. Other technologies, such as Handwriting Recognition seems not to have been reported as often as perhaps would have been expected, considering the increasing popularity of devices such as PDAs, Tablet Notebooks and the Nintendo DS, which utilize the technology. Often the issue here is the transparency of the technology; many users of SALT simply do not realise that the devices they employ use SALT technology to accomplish various functions.

The most regularly[29] used SALT technologies amongst respondents were (in order of popularity):

- Text proofing tools (spelling/grammar checkers)
- Electronic language resources (such as online dictionaries)
- OCR (optical character recognition) software
- Computer-aided translation (translation memory) software
- Other language analysis software
- Machine translation software
- Text-to-speech software
- Speech recognition software

Below are tables listing the results for each type of SALT individually:

[29] 'Regularl use' being defined as once a month or more frequently

How often do you use	Speech recognition
Once a day or more often	3
At least once a week	2
At least once a month	1
At least once every six months	1
Less often than once every six months	1
Never	20
Don't know	0
	28

How often do you use	Speaker recognition
Once a day or more often	1
At least once a week	0
At least once a month	0
At least once every six months	0
Less often than once every six months	0
Never	27
Don't know	0
	28

How often do you use	Optical character recognition
Once a day or more often	2
At least once a week	2
At least once a month	6
At least once every six months	4
Less often than once every six months	6
Never	8
Don't know	0
	28

How often do you use	Handwriting recognition
Once a day or more often	1
At least once a week	0
At least once a month	3
At least once every six months	2
Less often than once every six months	4
Never	18
Don't know	0
	28

How often do you use	Spelling and/or grammar checkers
Once a day or more often	21
At least once a week	4
At least once a month	2
At least once every six months	0
Less often than once every six months	0
Never	0
Don't know	1
	28

How often do you use	Other language analysis apart from spelling and/or grammar checking
Once a day or more often	4
At least once a week	3
At least once a month	2
At least once every six months	2
Less often than once every six months	2
Never	14
Don't know	1
	28

How often do you use	Automatic generation of language by computer (grammars that create text sentences)
Once a day or more often	0
At least once a week	0
At least once a month	0
At least once every six months	1
Less often than once every six months	2
Never	23
Don't know	2
	28

How often do you use	Text-to-speech systems
Once a day or more often	2
At least once a week	1
At least once a month	4
At least once every six months	1
Less often than once every six months	1
Never	17
Don't know	2
	28

How often do you use	Dialogue modelling (computer models for analysing human dialogue)
Once a day or more often	1
At least once a week	2
At least once a month	0
At least once every six months	1
Less often than once every six months	2
Never	22
Don't know	0
	28

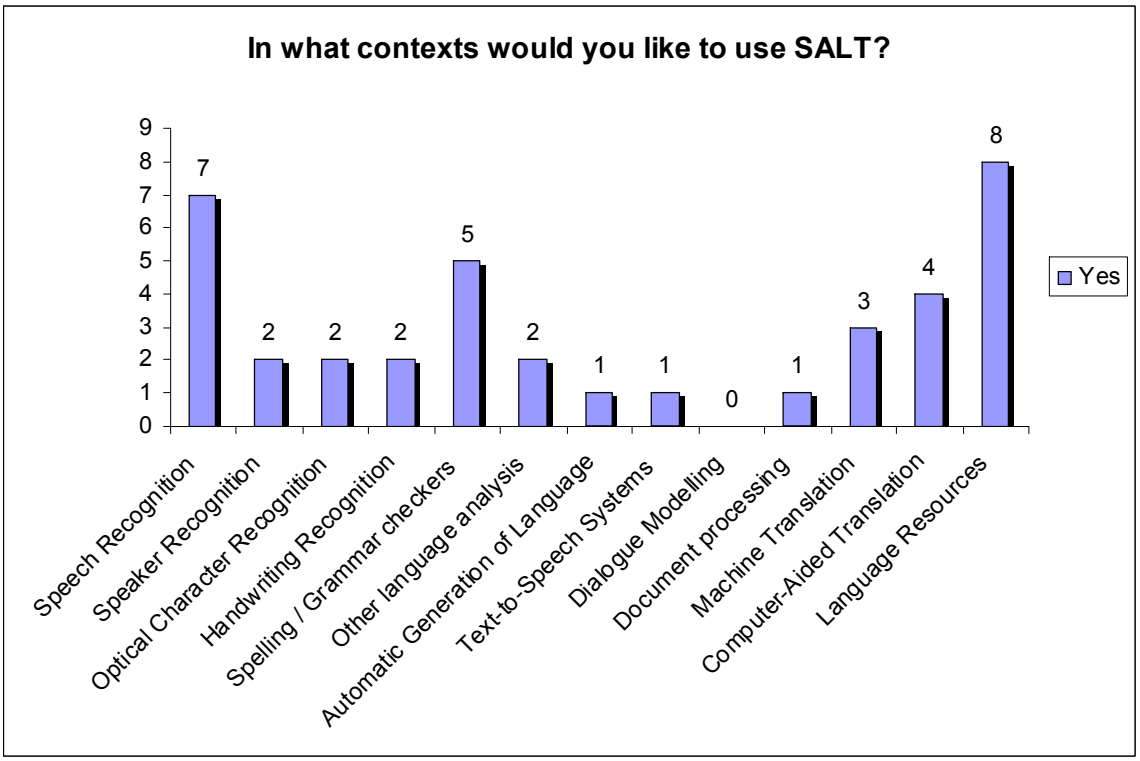
How often do you use	Document processing
Once a day or more often	2
At least once a week	1
At least once a month	0
At least once every six months	2
Less often than once every six months	0
Never	21
Don't know	2
	28

How often do you use	Machine translation
Once a day or more often	1
At least once a week	0
At least once a month	7
At least once every six months	4
Less often than once every six months	0
Never	16
Don't know	0
	28

How often do you use	Computer-aided translation
Once a day or more often	4
At least once a week	5
At least once a month	1
At least once every six months	0
Less often than once every six months	1
Never	15
Don't know	2
	28

How often do you use	Language resources
Once a day or more often	15
At least once a week	5
At least once a month	4
At least once every six months	1
Less often than once every six months	0
Never	3
Don't know	0
	28

8. (0011: E1.) Prospective Salt Users - In what context or contexts would you like to use SALT?



Respondents who had indicated that they did not use SALT technologies (but would like to) were asked to specify what were the kind of SALT technologies they would like to use.

Language Resources such as online dictionaries proved most popular, reflecting the importance of these writing aids to users in general writing tasks. Closely following Language Resources in terms of popularity was the category of Speech Recognition, as it is widely seen as possessing the potential to allow for a more efficient means for people to interface with technology.

Spelling and Grammar Checkers were in third place, and, as with Language Resources, their popularity reflects the continued importance of the written word, and the desire to communicate without having spelling and grammar errors dilute the strength of the message.

Next in terms of popularity were *Computer-aided Translation* and *Machine Translation*. Their popularity reveals the importance of translation both in bilingual Wales and in the wider multilingual world, and demonstrates an understanding amongst some respondents of the capacity of technology to increase the quality and efficiency of translations produce by the translation industry in Wales.

Speaker Recognition, *Optical Character Recognition*, *Handwriting Recognition* and *Other Language Analysis* followed, and all received an equal amount of responses. These are more specialized technologies which are employed in more specific circumstances, as in the case of *OCR* or *Speaker Recognition*. Although OCR is well established (at least in the major languages) the other technologies are becoming more and more familiar as the devices which use these technologies become more available. At the lower end of the popularity scale are more developer orientated technologies such as Automatic Generation of Languages, and Document Processing each received a single response. Their position here reflects the fact that they are mostly employed by specialized SALT developers working in specific categories, to whom they are vitally important. Text-to-Speech Systems also figured amongst the less popular technologies, but this in part reflects lack of responses from users with visual impairments, who are the primary users of the technology, and who cannot access text based technology such as the internet without it.

Below are the results by SALT Category in table format:

Would you like to use :	**Speech Recognition**
Yes	7
No	2
	9

Would you like to use :	**Speaker Recognition**
Yes	2
No	7
	9

Would you like to use :	**Optical Character Recognition**
Yes	2
No	7
	9

Would you like to use :	**Handwriting Recognition**
Yes	2
No	7
	9

Would you like to use :	**Spelling / Grammar checkers**
Yes	5
No	4

	9

Would you like to use :	**Other language analysis**
Yes	2
No	7
	9

Would you like to use :	**Automatic Generation of Language**
Yes	1
No	8
	9

Would you like to use :	**Text-to-Speech Systems**
Yes	1
No	8
	9

Would you like to use :	**Document processing**
Yes	1
No	8
	9

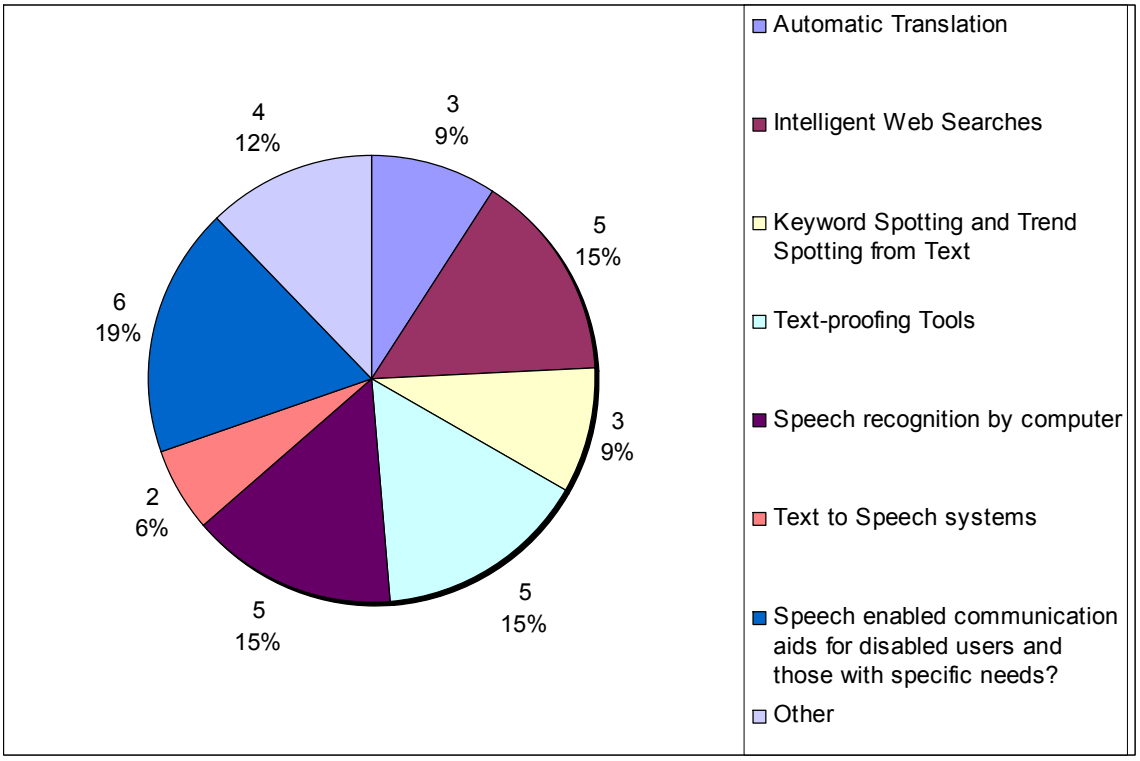
Would you like to use :	**Machine Translation**
Yes	3
No	6
	9

Would you like to use :	**Computer-Aided Translation**
Yes	4
No	5
	9

Would you like to use :	Language Resources
Yes	8
No	1
	9

9. (0012: F1.) SALT Developers - What SALT products or techniques do you currently develop, or have you developed in the past?

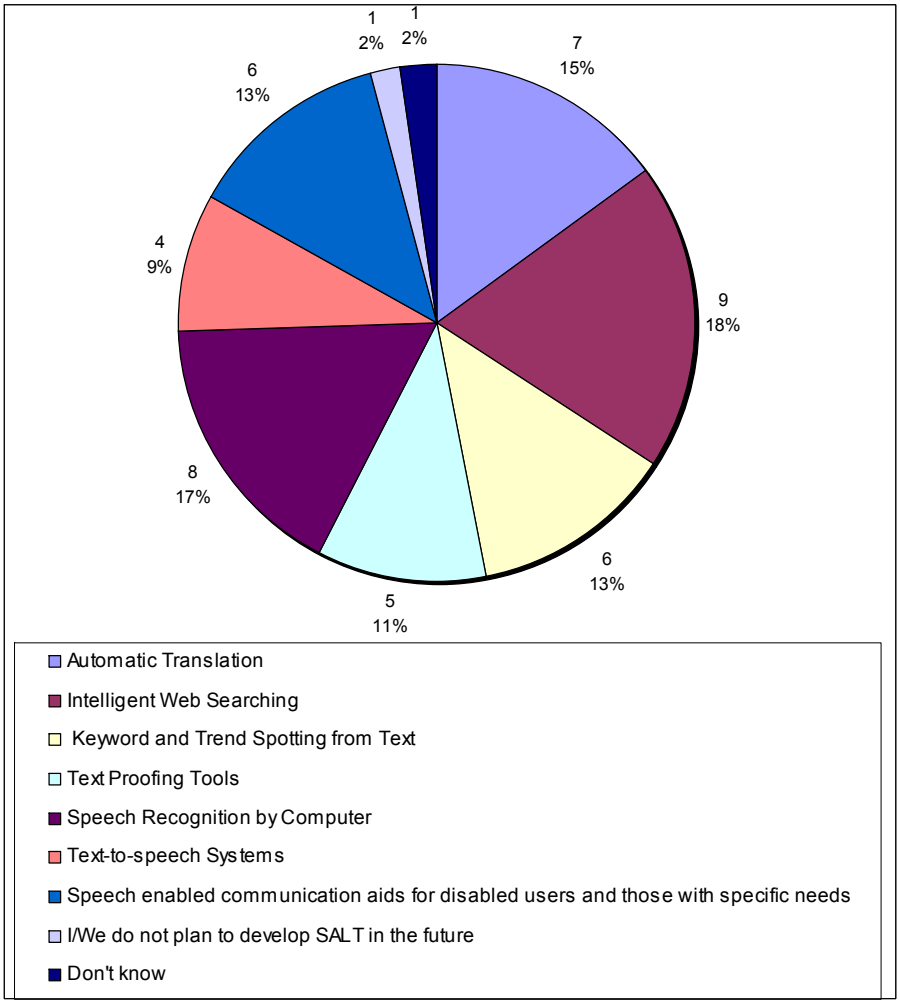
Results for: SALT Developers (19 of 48 respondents)
Of the 19 SALT Developers, 6 (19%) indicated that they developed *Speech Enabled Communication Aids for Disabled Users and those with Specific Needs*, 5 (15%) indicated that they developed *Text-proofing Tools*, 5 (15%) indicated that they developed *Speech Recognition by Computer*, 5 (15%) indicated that they developed *Intelligent Web Searches*, 3 (9%) indicated that they developed *Automatic Translation*, 3 (9%) indicated that they developed *Keyword Spotting and Trend Spotting from Text*, 2 (6%) indicated that they developed *Text-to-Speech Systems*, and 4 (12%) indicated that they developed SALT types other than those listed.



3	5	3	5	5	2	6
Automatic Translation	**Intelligent Web Searches**	**Keyword Spotting and Trend Spotting from Text**	**Text-proofing Tools**	**Speech recognition by computer**	**Text to Speech systems**	**Speech enabled communication aids for disabled users and those with specific needs?**

10. (0013: F2.) SALT Developers - What SALT do you intend to develop in the future?

Results for: SALT Developers (19 of 48 respondents)
When asked to specify the SALT they intended to develop in the future, the 19 SALT Developers the responded as follows: 9 (18%) indicated that they would *develop Intelligent Web Searches*, 8 (17%) indicated would develop *Speech Recognition by Computer*, 7 (15%) indicated that would develop *Automatic Translation*, 6 (13%) indicated would develop *Speech Enabled Communication Tools for Disabled Users and those with Specific Needs*, 6 (13%) indicated would develop *Keyword Spotting and Trend Spotting from Text*, 5 (11%) indicated they would develop Text-proofing Tools, 4 (9%) indicated they would develop *Text-to-Speech Systems*, and 1 (2%) each indicated that they were not intending to develop SALT in the future, or didn't know.

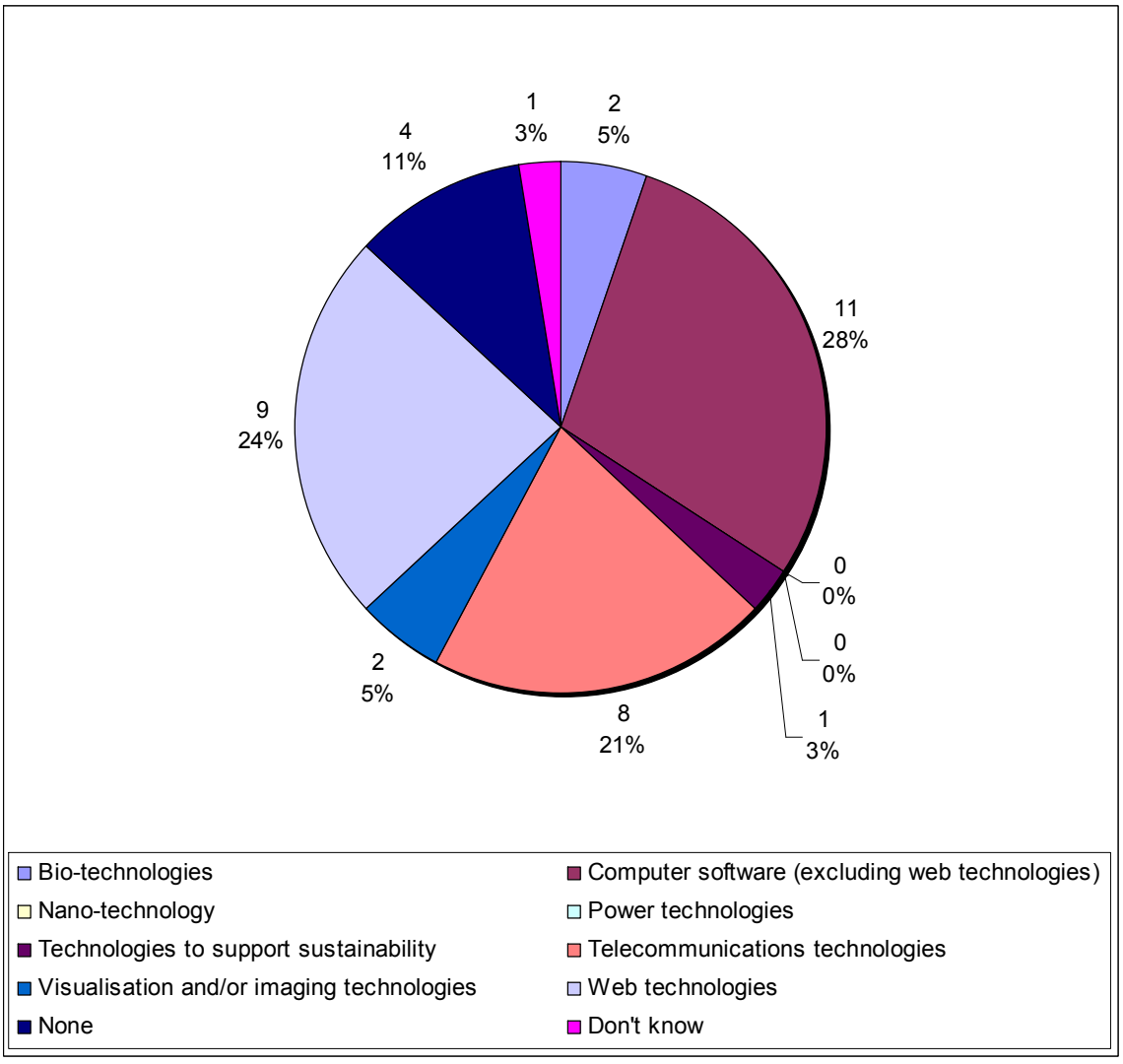


	Automatic Translation	Intelligent Web Searching	Keyword and Trend Spotting from Text	Text Proofing Tools	Speech Recognition by Computer	Text-to-speech Systems	Speech enabled communication aids for disabled users and those with specific needs	I/We do not plan to develop SALT in the future	Don't know
Yes	7	9	6	5	8	4	6	1	1
No	12	10	13	14	11	15	13	18	18

Intelligent Web Searching's popularity reflects the need to improve the results of web based searching due to the enormous amounts of information to be found on the web. Current search methods concentrate mostly on ranking results by popularity, rather than by their relevance to the context of the user's search. Highly inflected languages such as Welsh would also benefit greatly from more intelligent searches which could identify various forms of the same word as being relevant to a search for that word. Automatic Translation also figures prominently in the results, being a double priority in a bilingual nation that is part of a wider multilingual world. Text Proofing Tools, despite their maturity as a technology, are also indicated b many to be the subject of further development. Text-to-speech synthesis, primarily of Welsh, is also another category where future development is intended so as to address a current imbalance between English and Welsh provision of the technology.

11. (0014: F3.) Developers of Salt - What other technologies do you develop?

Results for: SALT Developers (19 of 48 respondents)
Of the 19 SALT Developers, 11 (28%) indicated that they also developed *Computer Software*, 9 (24%) indicated that they developed *Web Technologies*, 8 (21%) indicated that they developed *Telecommunication Technologies*, 4 (11%) indicated that they developed no other technologies apart from SALT, 2 (5%) indicated that they developed *Visualization Technologies*, 2 (5%) indicated that they developed *Bio-technologies*, 1 (3%) indicated they developed *Technologies to Support Sustainability* and 1 (3%) indicated that did not know what other technologies they developed.



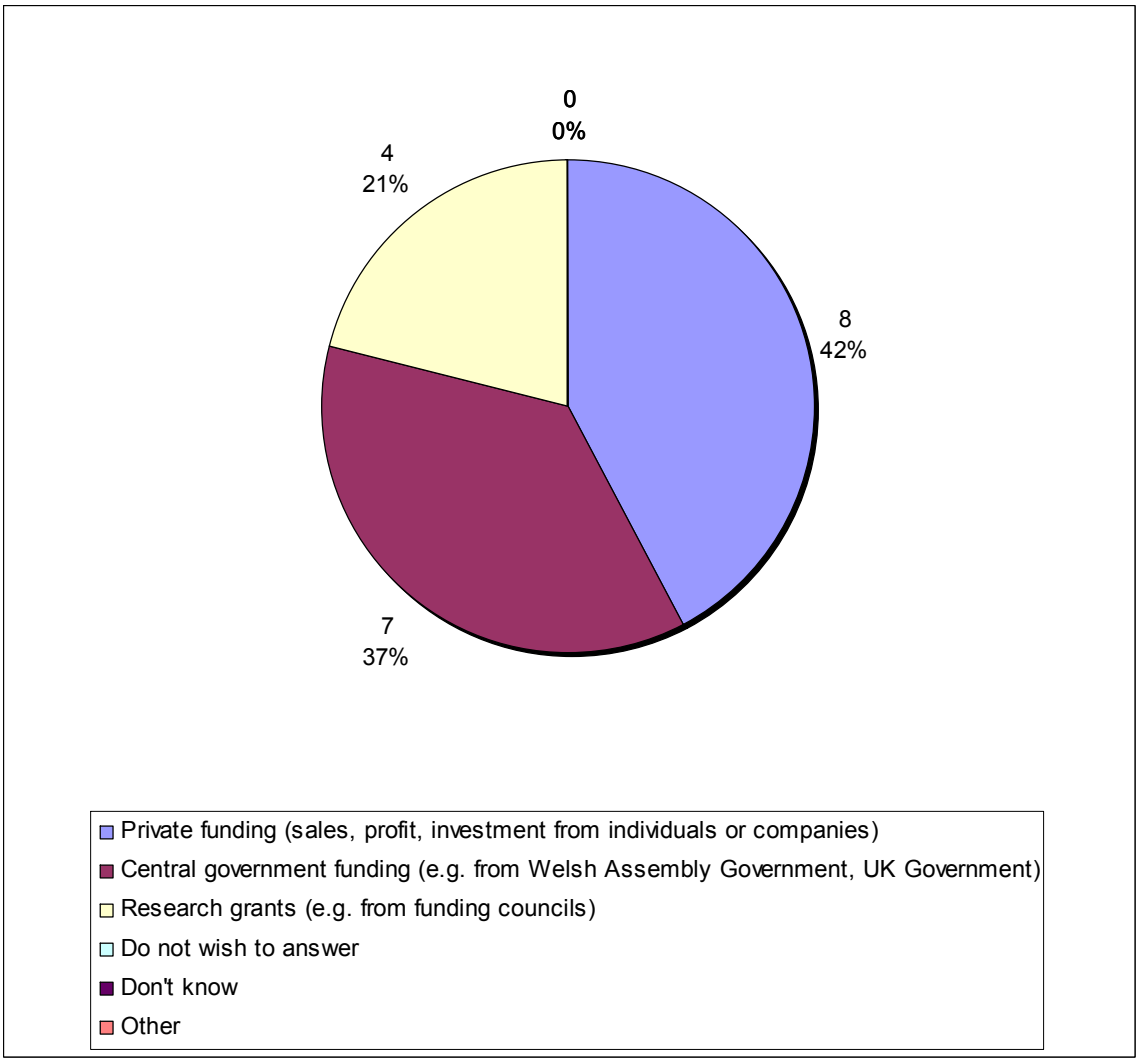
Bio-technologies	Computer software (excluding web technologies)	Nano-technology	Power technologies	Technologies to support sustainability	Tele-communications technologies	Visualisation and/or imaging technologies	Web technologies	None	Don't know
2	11	0	0	1	8	2	9	4	1

Interestingly, only 4 (11%) of those who identified themselves as **SALT Developers** were purely SALT Developers. The vast majority of respondents claiming to be SALT

developers also developed other technologies, and may in fact be primarily developers of these technologies. The largest proportion of these developers developed Computer Software, closely followed by Web Technologies, reflecting the interdisciplinary nature of SALT and the degree that skills from software and web development are transferable to the field of SALT. However, many aspects of SALT are highly specialized, and in many respects these results could indicate that many of the respondents who identified themselves as SALT developers are in fact implementing SALT components in their work rather than developing SALT technologies. SALT's importance to telecommunications is highlighted by the high proportion of responses received from parties involved in telecommunication development, and reflects its use with automated telephone services such as cinema ticket booking lines and automated flood alerts by telephone.

12. (0015: F4.) Developers of SALT - What is your main source of funding?

Results for: SALT Developers (19 of 48 respondents)
Of the 19 SALT Developers, 8 (42%) identified *Private Funding* as their main source of funding, 7 (37%) identified *Central Government Funding* as their main source of funding, and 4 (21%) identified *Research Grants* as their main source of funding.



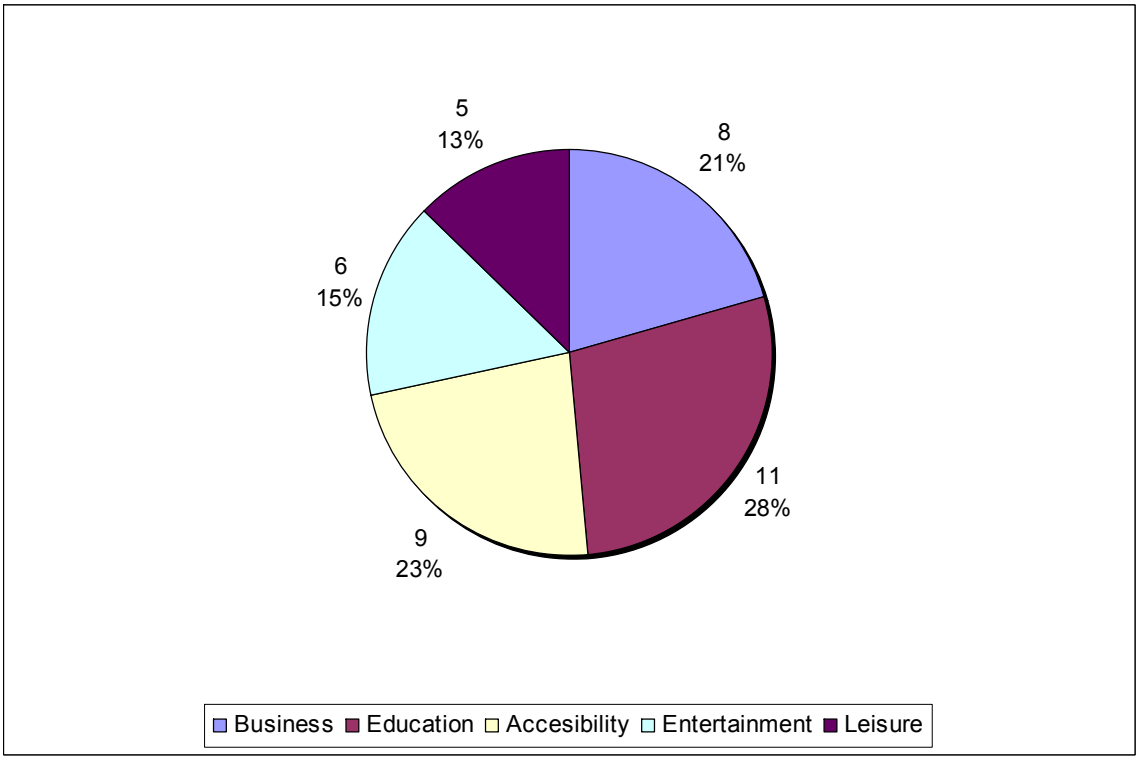
Private funding (sales, profit, investment from individuals or companies)	8
Central government funding (e.g. from Welsh Assembly Government, UK Government)	7
Research grants (e.g. from funding councils)	4
Do not wish to answer	0
Don't know	0
Other	0
	19

Three main sources of funding for respondents identifying themselves as SALT developers were identifiable from the survey results. The largest proportion of respondents who identified themselves as SALT developers were financed mainly by private funding. The next largest proportion was financed mainly by Central Government

Funding, and the final proportion responded that they were funded mainly by Research Grants. However, note this does not reflect the amount of money involved in SALT development – a single research grant awarded to a Higher Education establishment may be larger than the Private Funding secured by many separate respondents. Also, due to the fact that many of those identifying themselves as SALT developers are not primarily SALT developers, in many cases a large proportion of this funding may not be being spent on SALT research at all.

13. (0016: F5.) At what markets do you target the SALT that you develop?

Results for: SALT Developers (19 of 48 respondents)
The 19 SALT Developers targeted the SALT they produce at the following markets: 11 (28%) targeted *Education*, 9 (23%) targeted *Accessibility*, 8 (21%) targeted *Business*, 6 (15%) targeted *Entertainment*, and 5 (13%) targeted *Leisure*.

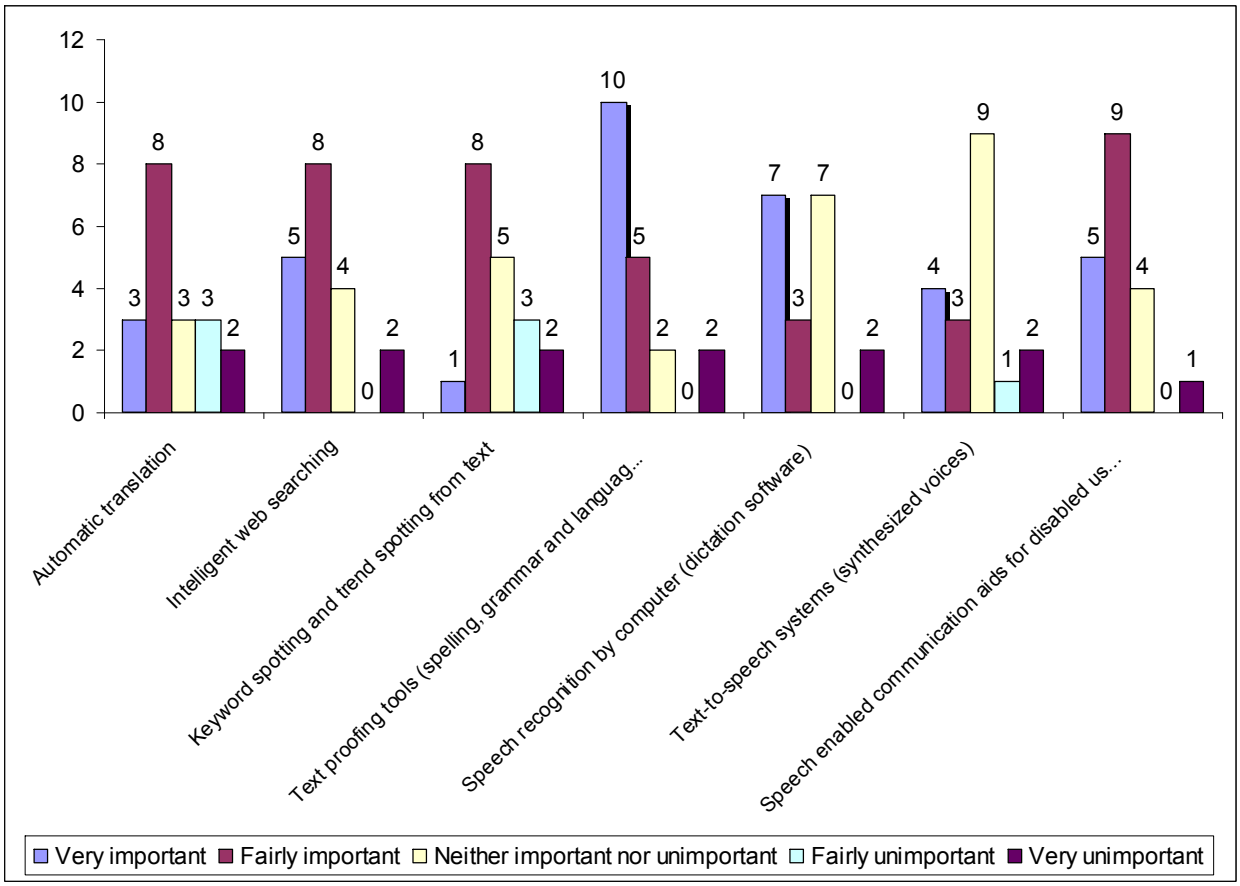


At what markets do you target the SALT that you develop?	Business	Education	Accessibility	Entertainment	Leisure
Yes	8	11	9	6	5
No	11	8	10	13	14
	19	**19**	**19**	**19**	**19**

Education, *Accessibility* and *Business* are the main markets targeted by those claiming to develop SALT, reflecting the fact that SALT technologies are mainly enabling technologies. However, in many ways these markets are intertwined, with accessibility important to business, and businesses specializing in education. SALT technology is therefore suitable for many markets, even those of Leisure and Entertainment, as evidenced by the example of Cineworld's speech recognition enabled ticket line for its cinemas, which include those in Llandudno, Newport and Cardiff.

14. (0017: F6.) SALT Developers - How important are the following aspects of SALT to you?

Results for: SALT Developers (19 of 48 respondents)

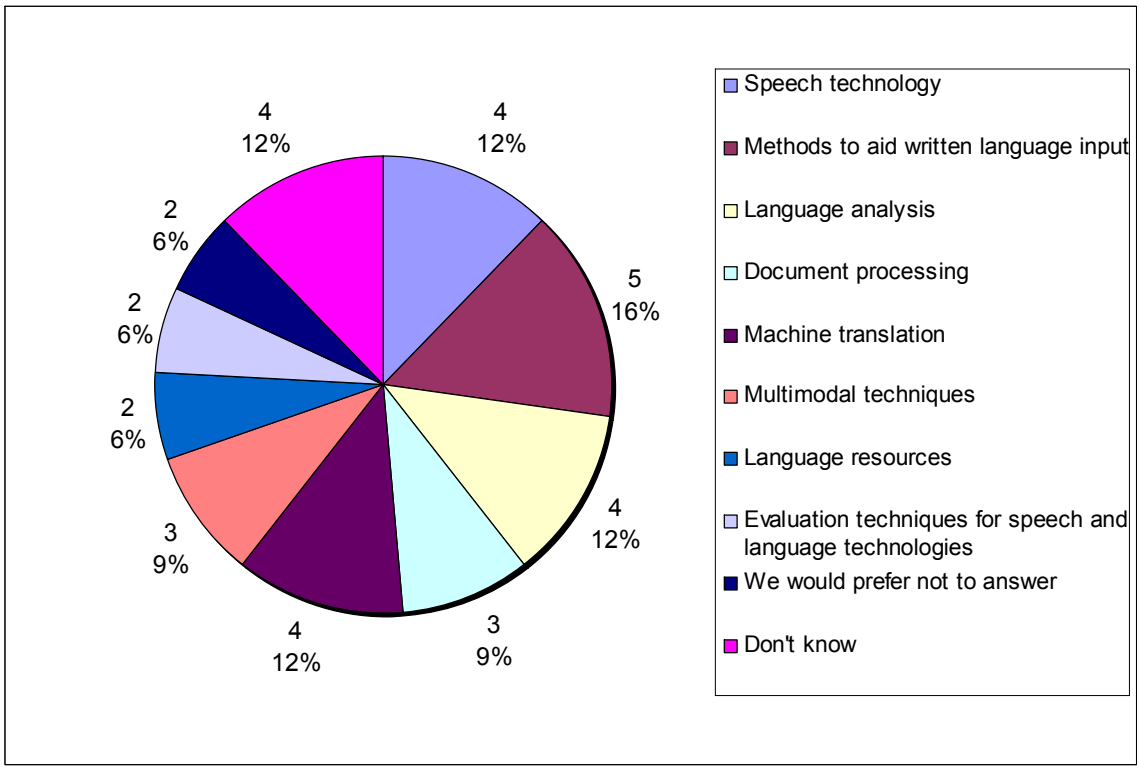


The results above show the importance of text-proofing tools to many of the developers who responded to the survey. As we have seen above, these technologies are widely used on a daily basis and are important to users in general. For SALT developers, many of the building blocks developed in the creation of text-proofing tools are invaluable for use in more specialized SALT technologies such as text-to-speech. As a whole, the results show that SALT developers consider a wide variety of SALT categories to be important, with very few considering a type of SALT to be very unimportant.

15. (0018: G1.) Prospective SALT developers - What SALT are you interested in developing in the future?

Results for: Prospective SALT Developers (14 of 48 respondents)

Of the 14 survey respondents who indicated that they were **Prospective SALT Developers**, 5 (16%) stated an interest in developing *Methods to Aid Written Language Input* , 4 (12%) stated an interest in developing *Language Analysis Tools* , 4 (12%) stated an interest in developing *Speech Technology*, 4 (12%) stated an interest in developing *Machine Translation*, 3 (9%) stated an interest in developing *Multimodal Techniques*, 2 (6%) stated an interest in developing *Language Resources*, 2 (6%) stated an interest in developing *Evaluation Techniques for SALT*, 2 (6%) did not wish to answer, and 4 (12%) did not know.

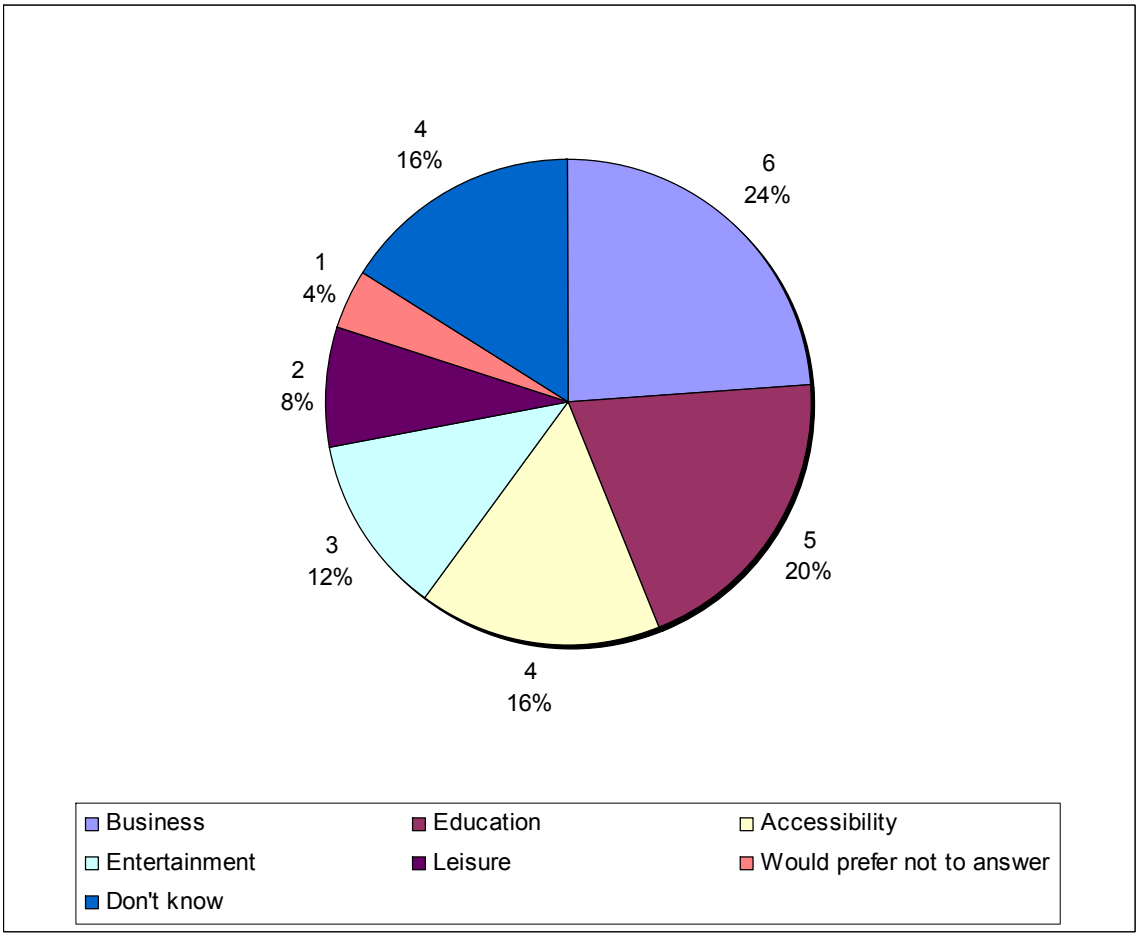


Speech technology	Methods to aid written language input	Language analysis, understanding and generation	Document processing	Machine translation	Multimodal techniques	Language resources	Evaluation techniques for speech and language technologies	We would prefer not to answer	Don't know
4	5	4	3	4	3	2	2	2	4
10	9	10	11	10	11	12	12	12	10
14	**14**	**14**	**14**	**14**	**14**	**14**	**14**	**14**	**14**

16. (0019: G2.) Prospective SALT developers - At what markets would you target the SALT that you would develop?

Results for: Prospective SALT Developers (14 of 48 respondents)

The 14 Prospective SALT Developers indicated that they would target the SALT they would develop at the following markets: 6 (24%) would target *Business*, 5 (20%) would target *Education*, 4 (16%) would target *Accessibility*, 3 (12%) would target *Entertainment*, 2 (8%) would target *Leisure*, 1 (4%) preferred not to answer and 4 (16%) did not know.



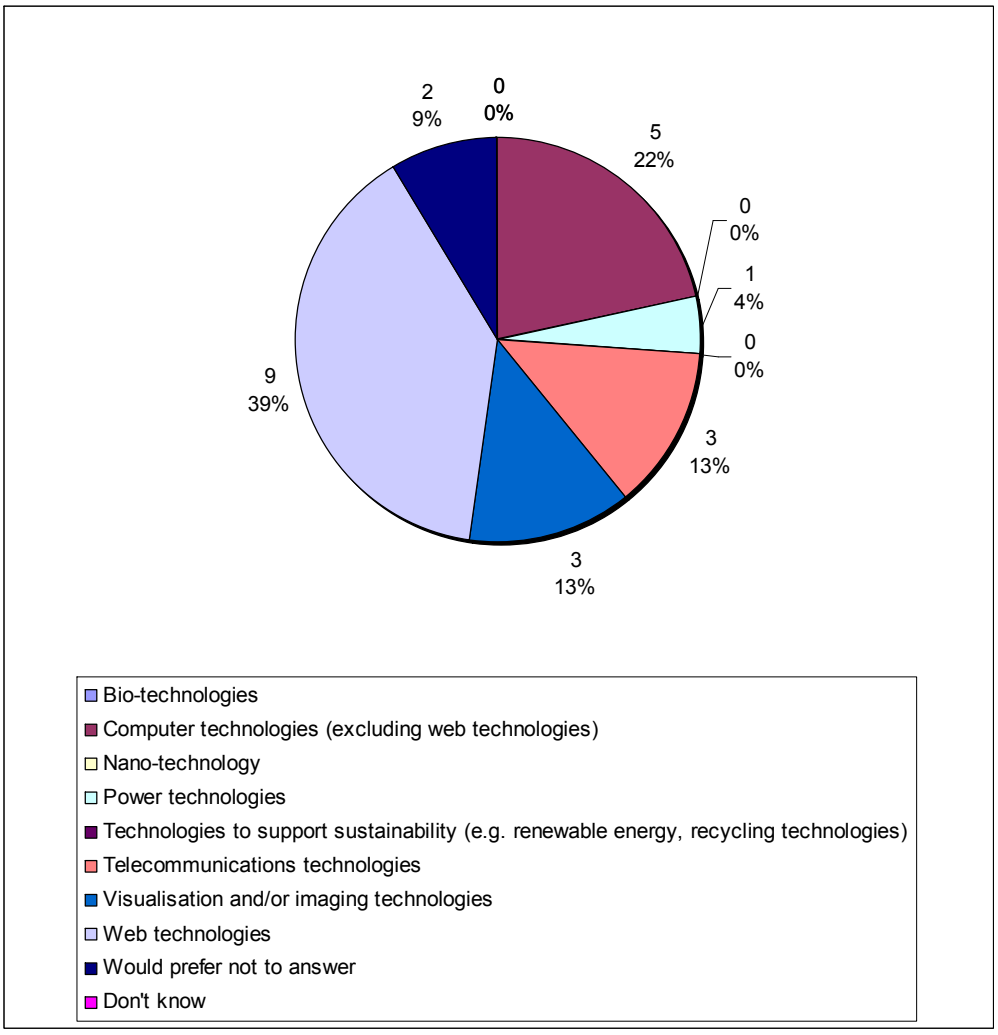
Business	Education	Accessibility	Entertainment	Leisure	Would prefer not to answer	Don't know
6	5	4	3	2	1	4

The results for Prospective SALT developers differ from those of SALT Developers in that Business is considered to be the main market that they would target, whereas Business was ranked third by SALT Developers. However, similarity of the number of responses for each category seems to indicate that potential SALT developers attempt to market their SALT to any available market.

17. (0020: G3.) What other technologies do you currently develop?

Results for: Prospective SALT Developers (14 of 48 respondents)

Of the 14 survey respondents who indicated that they were **Prospective SALT Developers**, 5 (22%) indicated that they also developed *Computer Software*, 9 (39%) indicated that they developed *Web Technologies*, 3 (13%) indicated that they developed *Telecommunication Technologies*, 3 (13%) indicated that they developed *Visualization and/or Imaging Technologies*, 1 (4%) indicated that they developed *Power Technologies*, and 2 (9%) did not wish to answer.



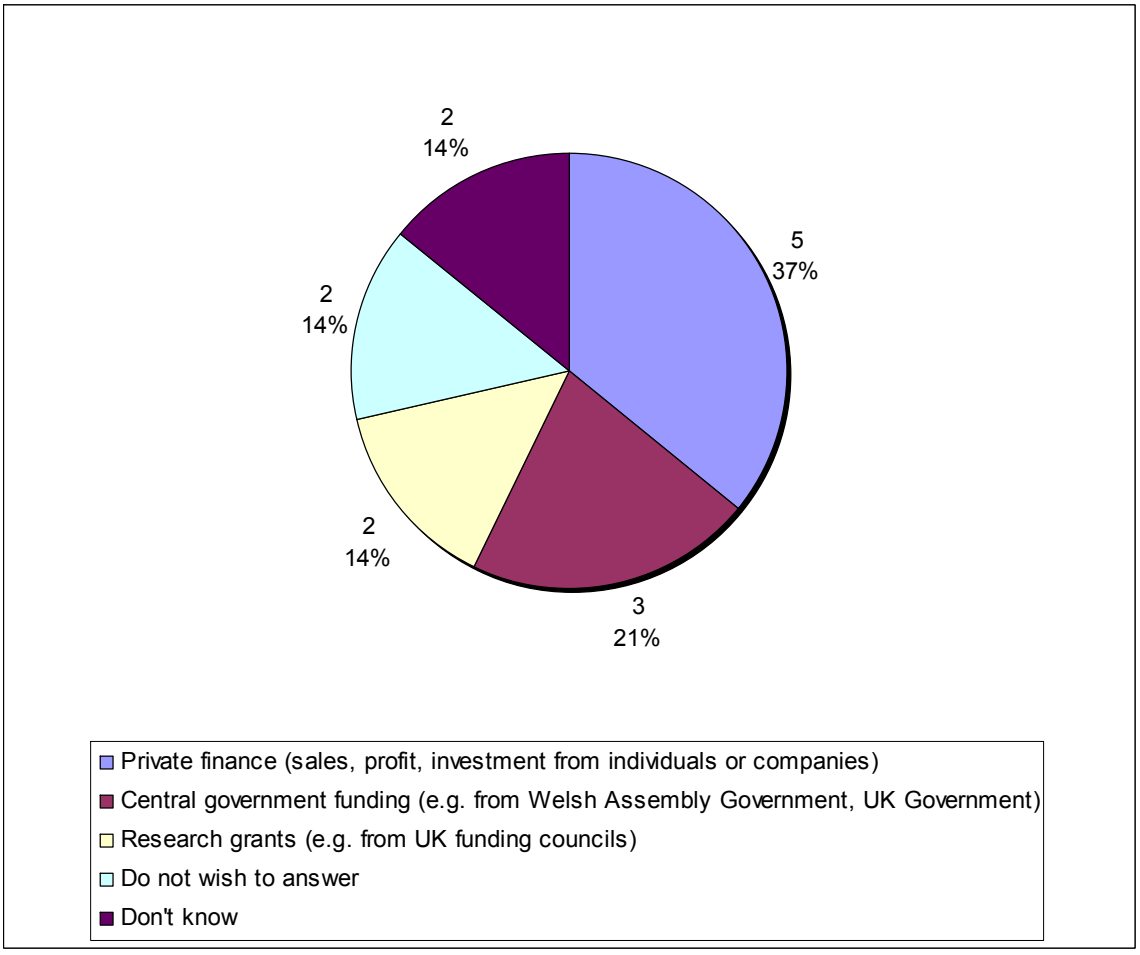
Bio-technologies	Computer technologies	Nano-technology	Power technologies	Technologies to support sustainability	Telecommunications technologies	Visualisation and/or imaging technologies	Web technologies	Would prefer not to answer	Don't know
0	5	0	1	0	3	3	9	2	0
14	9	14	13	14	11	11	5	12	14
14	**14**	**14**	**14**	**14**	**14**	**14**	**14**	**14**	**14**

Similarly to the case of SALT Developers, the largest proportion of Prospective SALT Developers develop *Web Technologies* and *Computer Software*, with a significant proportion from the *Telecommunication* and *Visualization* sector. Again, this reflects the interdisciplinary nature of SALT and the degree that skills from software and web development are transferable to the field of SALT. However, it could be that these Prospective SALT Developers in fact intend to implement SALT components in their work rather than develop SALT technologies of their own.

18. (0021: G4.) What is your main source of funding?

Results for: Prospective SALT Developers (14 of 48 respondents)

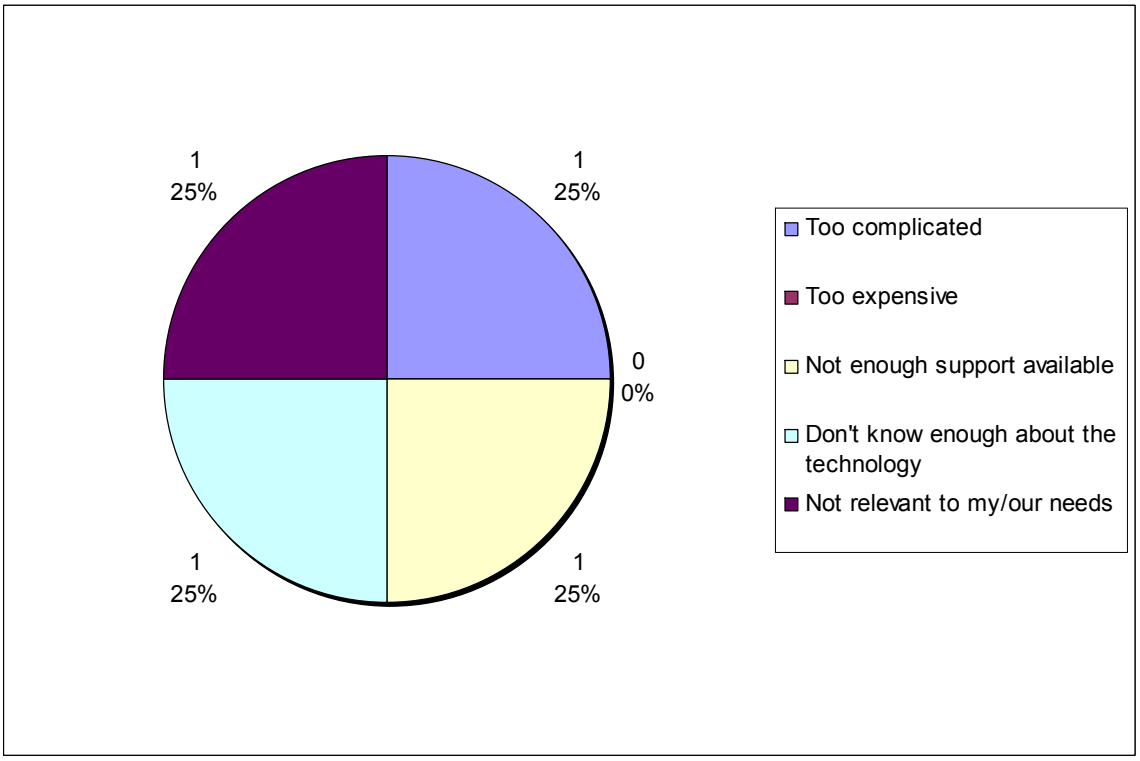
Of the 14 survey respondents who indicated that they were **Prospective SALT Developers**, 7 (50%) were *Private Individuals*, 4 (29%) were from *Companies or Commercial Organizations*, and 2 (14%) were *Higher Education Establishments*. One respondent (7%) replied from *Health or Care Establishment*, *Not-for-profit Organization or Charity* and *Other*.



Private finance (sales, profit, investment from individuals or companies)	5
Central government funding (e.g. from Welsh Assembly Government, UK Government)	3
Research grants (e.g. from UK funding councils)	2
Do not wish to answer	2
Don't know	2
	14

As in the case of SALT Developers, the largest proportion of respondents who identified themselves as SALT developers were financed mainly by private funding, with a higher percentage of respondents than for the SALT Developers. Again, the next largest proportion was financed mainly by Central Government Funding, and the final proportion responded that they were funded mainly by Research Grants.

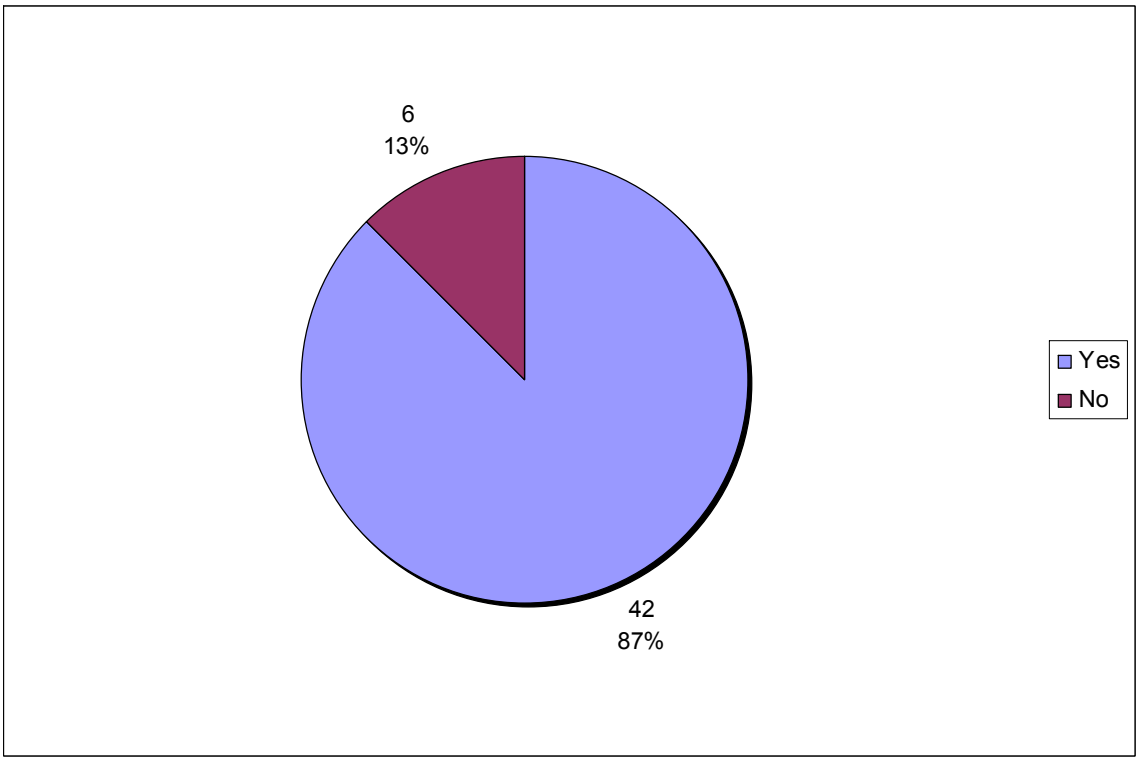
19. (0022: H1.) Reasons (for no interest in SALT):



Too complicated	Too expensive	Not enough support available	Don't know enough about the technology	Not relevant to my/our needs
1	0	1	1	1
3	**3**	**3**	**3**	**3**

Only three respondents indicated that SALT was of no interest to them, with different reasons given by each respondent (one respondent gave two reasons). In the free text box, one respondent claimed that there was no SALT technology available in Welsh, despite this not being the case. It would seem that this is another example of confusion about what exactly constitutes SALT.

20. (0029) Are you happy to be contacted with further information about the SALT Cymru project?



Yes	42
No	6
	48

A measure of the interest in SALT and the SALT Cymru project is that 87% of those who completed the survey were happy to receive further information about further developments, a result which bodes well for the establishment of special interest groups.

Key findings from survey

There was a pleasing interest shown towards SALT in Wales with a higher than expected number of respondents, the vast majority of which showed some degree of interest in SALT, whether as developers, prospective developers or as users. The awareness of SALT and perceptions of SALT varied considerably amongst survey respondents. This might not have expected to have been the case; the survey group, comprised as they were of people who had opted to complete a questionnaire in this technical field, would reasonably have been expected to be self-selecting and aware of the areas in which they would be answering questions. However, despite comprehensive and helpful definitions alongside survey questions and on the SALT Cymru website, there was still considerable confusion over what constituted SALT. In particular, some respondents to the survey did not consider themselves to be SALT users, yet it can be reasonably expected that they would use spelling and grammar checkers (considered part of SALT according to the definition) on a regular basis. Some technologies such as Handwriting recognition may have received less interest due to not realising their use handheld devices such as PDAs and the Nintendo DS, whilst Intelligent Web Searching may have received a disproportionate amount of interest due to what its name represented in the mind of the respondent rather than the reality of its proposed benefits in comparison to traditional web searches. While SALT Cymru is not directly concerned with educating users in what SALT is, the issue of how to communicate to users what their SALT needs might be should be borne in mind. It is discussed in more detail in Section 23, which considers the role of the special interest group.

Those who responded to the survey as SALT Developers seem to belong to two separate categories: those developing SALT components, and those who implement those SALT components in their products. The first category consists almost exclusively of developers from, or connected to, Higher Education. The second category consists of developers who primarily work with web technologies, computer software and telecommunications, but who need to integrate SALT components into their products to cater for the accessibility, language or educational needs of their users. This second category may largely account for the 76% of SALT developers who stated that they also developed other technologies within their organisation, although the interdisciplinary nature of SALT and the transferability of SALT skills to software and web development are also valid reasons.

Higher Education performs both the role of developing of SALT technology and its subsequent implementation. Despite this apparent strength in academia, the survey shows that Higher Education development is dependent on research grants, making its future highly vulnerable to fluctuating fortunes in grant capture. This means that a barren period could wipe out most of the SALT research base in Wales. This in turn would affect those attempting to implement or commercialize their SALT output.

Many of those who indicated an interest in developing SALT in the future seemed to be micro SMEs involved in developing software and web technologies which would benefit from including SALT technology. This was especially true of developers who were not SALT specialists but who used SALT components in their products, e.g. to improve accessibility to disabled users. These need to be catered for with imaginative ways of

bridging the gap between pre-competitive research provided by universities and useful low-cost components that they could include in their products.

The influence of catering for the needs of the Welsh-language on SALT development in Wales is demonstrated by the high incident of bilingual websites possessed by developers and by the high incident of text-proofing tools, language resources that the results indicated were being developed, most of which are products specific to the Welsh language. The survey results as a whole demonstrates the need for these SALT technologies, and that the capacity to service this need does exist, although no indication as to the sustainability of this situation is provided. Currently however, there is a demonstrable home-grown expertise in the field of SALT which could be extended to other languages in order to exploit markets beyond Wales. The development of translation technologies, again deriving from the need to cater to the bilingual nature of Wales was strongly indicated as an avenue of future development by SALT developers. Developing and improving such translation technologies could provide considerable cost savings for public administration in Wales, and commercial exploitation possibilities in a global market.

As users of SALT, the Translation Industry is a significant sector which benefit from specific SALT tools, especially translation memory and computer aided translation tools. This survey did not go into further detail on the use of SALT by this sector, as it is the subject of a separate study by Tegau Andrews (PhD student at Bangor University). It is discussed elsewhere in this report.

While a wide variety of SALT areas were mentioned by respondents, some appeared to be of consistent interest to developers, potential developers and users. These included:

- Text-proofing tools
- Speech-enabled technologies (whether speech synthesis as a module, or speech-enabled communication aids as finished products)
- Speech recognition
- Intelligent web searching
- Keyword and trend spotting from text
- Machine translation

These were primarily aimed at aiding:

- Accessibility
- Education
- Welsh-language needs
- Communication needs
- Interfacing with technology

These provide a basis for a programme of work in SALT in Wales. Together with the encouraging response that 42 of the respondents would welcome further contact and information on the SALT Cymru project, we have the basis for the formation of special interest group, as well as a working list of SALT areas of interest to all.

Appendix E: Focus group discussions

E1: Association of Welsh Translators, Aberystwyth, November 16th

This focus group was held as part of the Association of Welsh Translators' annual meeting in Aberystwyth, 16th November 2007

Following a short presentation giving an overview of SALT in general, and of SALT Cymru's aims in particular, a lively focus group was held on translators' needs in SALT. It was decided to focus on specific technologies within the compass of SALT, and solicit opinions from those present on their current and future uses of those technologies.

About 80 people were present.

Speech recognition

At least two translators had used Dragon NaturallySpeaking or IBM ViaVoice in English. These are the two most widely used commercial speech recognition packages for the language. It was noted that such technologies were beneficial to those who were at risk of developing repetitive strain injury, and that it was good to be able to avoid typing for brief periods. It was stated that health and safety considerations for translators (in terms of avoiding constant typing) were a strong argument in favour of their using speech recognition systems.

It was also stated that Welsh translators normally undertook very little work in translating from Welsh into English, so that not many opportunities arose to use (English) speech recognition systems in their normal course of work. Additionally, some spoke of problems getting the Dragon system to recognise their accent, and it was noted that such systems did not always work effectively when the quality of the user's voice had changed (e.g. when they had a cold).

It was asked how many would use a Welsh-language speech recognition system were one to exist. Over half of those present showed an interest in this. It was also noted that this would benefit disabled translators who were unable to type. One participant suffered from a degenerative muscular condition which meant that he would be unable to use a conventional keyboard in a few years time. Speech recognition for Welsh would mean that he could continue to work as a translator where he would otherwise be unable to do so.

Translation memory software

The focus group leaders asked how many of the translators used translation memory software as part of their workflow. Slightly fewer than half those present (about 35 out of 80) used some sort of translation memory software. Of those who used such software, the majority (about 20) used Déjà Vu, about 10 used Trados, and about 5 used Wordfast.

However, no further questions were asked to reveal how many had the freedom to choose their own translation memory software, rather than be restricted to that which their supervisors had chosen centrally.

The majority agreed that they would be ready to use free translation memory software. It was requested that such software attempt to fulfil basic functionality effectively, rather than attempt to execute many functions less successfully. It was stated that it should be able to import many file formats, but for simplicity, should output in only one or two of those formats.

OCR (optical character recognition) software

This was used by some translators in order to scan documents before translating them. It was claimed that this worked adequately in welsh (with some exceptions, e.g. the digraph 'll' was sometimes detected as 'L1' by the software). It was noted that using Word's Welsh spell checker improved the quality of the OCR output.

Remote simultaneous translation

A small amount of interest was shown in this idea to create a 'virtual presence' for translators, enabling them to translate at meetings without physically attending them. The following points were raised:

- Are there similar ideas in the telecottages project formerly run by Antur Teifi?
- The Welsh Video Network connects further and higher education institutions in Wales: see http://www.wvn.ac.uk
- Is this practical in geographical areas which do not have access to broadband?

Other

It was noted that simple word counting software (which would also count material in text boxes), able to work with Microsoft Word (.doc) and Adobe Reader (.pdf) formats, would greatly benefit translators when attempting to charge for their work.

General

The point was raised that technologies, where appropriate, should appear on the largest possible number of platforms (including educational ones, e.g. Blackboard).

It was noted that the focus group were far more interested in certain technologies after they had been practically and clearly explained. It was essential, therefore, to produce materials explaining the technology in an informal, friendly way.

Appendix F. Accounts of the state of the art in particular areas of SALT

F1. Speech synthesis – the state of the art

There are an increasing number of scenarios where the desired output of a computer system must be speech-based. These may include:

- embedded systems where a visual display is not available
- control situations where a user could not keep their constant attention on a display
- automotive systems, where the user's attention must not be distracted from driving
- systems to enable blind or partially sighted users to access text-based information
- aids for users with dyslexia or other reading difficulties

Several different technologies may be used in speech synthesisers, which have different trade-offs in terms of quality of output, time taken for development, size of deployment and processing power required.

Most currently deployed speech synthesis systems use *concatenative synthesis* techniques, where pre-recorded utterances are segmented into words, or smaller units, and then concatenated together to produce the speech output. Pauses are added to the speech signal at points where phrases end in the input text. Intonation is included by modifying the speech signal using standard digital signal processing techniques.

A major challenge in speech synthesis is *tokenisation*, that of separating the input text into individual words and determining which words should be uttered. This is not a trivial problem, as abbreviations and groups of numbers should often be spoken differently depending on their context ('142 St. John St.', for example). If tokenisation is inadequate, many utterances may be spoken incorrectly, or not at all.

Diphone synthesis
A language consists of a number of unique speech sounds, termed *phonemes*. A phoneme represents the smallest unit of speech that can differentiate one word from another. In English, for example, the difference between the words 'dot' and 'dog' defines that one contains the phoneme /t/, and the other the phoneme /g/.

In early speech recognition experiments, it was thought that recording a language's individual phonemes, and concatenating them, would produce acceptable output. This, however, was found not to be the case; it is at phoneme boundaries that the greatest changes in speech characteristics are found. Simply concatenating the phonemes gives rise to a large number of disfluencies at phoneme boundaries, making much of the output unintelligible.

Diphone synthesis, in contrast, relies on concatenation of phoneme-to-phoneme transitions, resulting in much smoother output. Recordings take place of speakers uttering 'nonsense words', whose only function is to contain the specific diphones under question. These are then segmented, either by hand or, more commonly, by a speech synthesiser which has been set to detect the boundaries between phonemes. The result

is a database of diphones. The input speech, after tokenisation, is then passed through a phonetic lexicon which converts each word into its constituent phonemes. The required diphones are then concatenated and output, with appropriate intonation.

Diphone synthesis has the advantage of relatively low computational load, and (in terms of speech synthesis) relatively low disc space and memory requirements. The trade-off for this is in terms of reduced intelligibility compared to most other synthesis techniques. It should be noted, however, that users do familiarise themselves with speech synthesis voices after a period of use, so for those using speech synthesis for extended lengths of time, initial intelligibility is less of an issue than it otherwise might be.

Unit selection synthesis

Unit selection offers higher synthesis quality than diphone synthesis, though at greater computational expense and with higher memory and disc space requirements. Unit selection synthesis involves the concatenation of *units* of speech of varying sizes. In recording material for unit selection synthesis, a recording script is put together of common words in the language and common word fragments (such as syllables). Additionally, a full set of the language's diphones is usually recorded. As with diphone synthesis, these recorded items are segmented and stored in a database.

While playing back speech in unit selection synthesis, particular attention is paid to the selection algorithm, which determines which segments in the database will be concatenated together to produce the output speech. For highest quality synthesis, the segments should be as long as possible. Falling back to the diphone set is considered a last resort, as this significantly decreases the output quality.

It is less critical to add intonation to unit selection synthesis output than it is in diphone synthesis. If large units are taken and combined in the resynthesis process, a significant amount of the voice's original expressiveness is retained. As a result, it is common for unit selection synthesisers to be deployed with little or no additional intonation modules.

Limited domain synthesis

The diphone and unit selection systems described previously are designed to be able to speak any text that might be input. Consequently, the development time for both systems is long (particularly in the case of unit selection synthesis), and there may be variable quality in the output.

For many applications, only a limited number of utterances may be required. Examples of such applications include speaking clocks and telephony applications where the range of possible conversations is known.

Limited domain synthesis is essentially a specialised form of unit selection synthesis. The prompts for recording are carefully chosen, so that every section within the range of desired outputs has been covered.

Very high quality synthesis can be achieved using this method, as it involves the concatenation of long samples of speech. There should be little or no disfluency in the output, and in many cases the listener may be unaware that they are listening to synthesised speech, rather than utterances that have been recorded in full.

Formant synthesis

The synthesis methods previously described have relied on prompts being recorded, segmented and then resynthesized to produce the desired output. An alternative to the above methods is *formant synthesis.* Formant synthesis uses a model containing speech parameters such as its fundamental frequency, its quality and its noise level. These parameters are varied to produce the waveform which is the output speech.

Formant synthesis produces speech which is significantly less human-sounding than in concatenative synthesis methods. However, formant synthesis methods mean that the speech is less likely to contain disfluencies in the output (i.e. sounds 'smoother'), and hence can be replayed and understood at higher speaking rates. This benefits users of screen readers, where a large amount of text can be navigated quickly.

Formant synthesis also has the advantage of a small footprint, in terms of computational and memory requirements. It is thus particularly suited for hand-held devices such as mobile phones.

Current trends in speech synthesis

Much current work in speech synthesis involves optimising algorithms for unit selection. The work involves ensuring that the algorithms select as few segments as possible for a given utterance, thus increasing the average length of each segment used and hence reducing the risk of disfluency and increasing naturalness.

As the number of voice-enabled devices increases, speech synthesis also begins to appear on smaller devices than the standard desktop computer. Thus, work is also being carried out in porting speech synthesis to less powerful processors, with the consequent need to improve the efficiency of existing techniques and reduce their computational load.

Significant work is also being undertaken to improve intonation and phrasing models, in order to make the output speech sound more natural.

F2. Speech recognition – the state of the art

The ability to use a human voice to interact with a computer has long been a goal of research engineers. Research in the field commenced with initial experiments in the 1950s for a computer to detect a small number of words by decomposing them into the individual sounds of human speech. Nowadays speech recognition systems can be run on standard desktop computers and are included as components of operating systems. Speech recognition, therefore, may be said to have reached maturity, at least in English and commonly spoken languages of the world.

Speech recognition is a statistical process, which relies on probabilistic measures of the most likely sounds and words to be uttered at each point in human speech. The main challenge in speech recognition is thus the great variability of such speech. It changes not only between speakers but within the speech of individual speakers themselves: speech can change its characteristics according to its context, the speaker's environment, or their emotional state.

Most speech recognition systems consist of the following components:

- Front-end processing stage
- Acoustic modelling stage
- Phonetic lexicon
- Language modelling stage

Front-end processing
This stage in the recogniser takes input speech and converts it into a set of parameters which more efficiently discriminate between different sounds in human speech. This is typically done by converting the input from a waveform into its component frequencies, and performing an additional frequency analysis on that result. The parameters output from the front-end processing stage typically mimic the frequency response of the human ear.

Acoustic modelling stage
The acoustic modelling stage in the recogniser normally comprises a set of mathematical models. One model typically represents one of the phonemes (individual sounds) of speech in the language being recognised. This is not always the case, however, and in cases where the vocabulary of the system is small, whole words may be modelled instead. If phonemes are being modelled and sufficient training data is available, greater accuracy is often achieved by modelling them within the context of the sounds which follow and precede them (such modelling is termed triphone modelling).

It is evident that a significant amount of data is required to train the models in the acoustic modelling stage, in order to represent the range of speech that the recogniser must be able to detect. Any so-called *training data* must represent the intended use of the recogniser. It must, therefore, represent the typical accents, pitch ranges, environments and speaking styles that might be encountered in the recogniser's eventual use.

Many recognisers, including most of those used by the consumer as off-the-shelf packages for personal computers, include an adaptive training mode. In such modes, at

the same time as recognition is carried out, the user's speech is taken and used to adapt the existing models. The resulting models better reflect the user's speech, resulting in improved recogniser performance for that user. Other recognisers, such as those employed in telephony applications, may be retrained off-line, using similar algorithms, to improve their performance on the range of users that are commonly encountered.

Phonetic lexicon

In very small vocabulary speech recognition applications, the individual words to be recognised may be simply matched against each other. In most situations, however, the output of the acoustic modelling stage is in the form of a string of phonemes. For eventual text output, these must be converted to individual words. To this end, a *phonetic lexicon* is required, which consists of all the words the recogniser might be expected to recognise, together with one or more possible phonetic transcriptions for each. The lexicon is used to restrict the possible outputs of the recogniser in the recognition stage, which has the effect of increasing the accuracy of the phoneme recognition by reducing the large number of phonemes that may be output in combination.

Language modelling stage

To further improve the accuracy of recognition, and to make the eventual output more representative of human communication, a *language model* is commonly used. Typically, this consists of an *n-gram analysis* stage, which consists of a set of probabilities determining how likely one word is to follow another in a language. The number of words included in this analysis varies according to the complexity of the recognition task and the amount of processing power available to the application. Simple bigram analysis, where all that is taken into account is the probability of one word following another, requires a relatively small amount of computation. More sophisticated analysis, where possible sequences of four or more words may be considered, represents a greater load on the system's processor.

A large amount of data is required to train the language model, which is usually done through large text corpora of written material for the language to be recognised. Typically a corpus of over a million words is considered the minimum required for a bigram model: n-gram models dealing with larger sequences of words will require larger corpora. The corpora used need only be written, rather than spoken ones: language modelling deals entirely in the sequences of transcribed speech, rather than the process of transcribing the speech itself. The text contained in the corpus should ideally reflect the eventual use of the recognition application.

In large vocabulary applications, the language model may be adapted on-the-fly. This is accomplished in a manner similar to acoustic model adaptation. In language model adaptation, if the recogniser is sufficiently confident that it has correctly detected a sequence of words, it will adjust the probabilities within the language model to reflect this.

Dialogue-based speech recognition systems may also include a far more sophisticated natural language processing stage, where advanced grammars are used to determine what the user's query is. These are used in such applications as interactive voice response (IVR) systems for full or partial automation of call centres.

Current performance

The performance of speech recognition applications depend largely on their expected use, the range of speakers that may need to use them, and the complexity of the vocabulary that will be recognised. Simple 'yes/no' systems using a microphone headset, restricted to a small number of users, may achieve 99.5% or better accuracy.

For more sophisticated systems, the measure commonly used to assess performance is the *word error rate* (WER). The WER represents the percentage of words that are in error in a recogniser's output. A lower WER thus represents better recogniser performance. The expected WER depends on the complexity of the recognition task. For recognition of spontaneous speech uttered over a telephone line, a WER of under 20% is common. Transcription of broadcast television programmes currently represents one of the most challenging areas for speech recognition. In this field, a WER of about 30% may be achieved for magazine programmes, where the breadth of topics, and hence the language used, may be very wide-ranging. By comparison, transcription of a weather report, where both the vocabulary and language patterns are restricted, may achieve a WER under 10%.

Current trends in speech recognition research

Current topics in speech recognition research include the investigation of multilingual acoustic modelling, where phonemes that are mapped in multiple languages may be modelled together. This represents a decrease in the amount of training data required for multilingual systems, as material that is common between languages may be shared between them, rather than being restricted to training one language's models alone. It also represents a way to reduce the amount of effort required to develop a new language or application for speech recognition, as training data may be 'borrowed' from a similar language for which a large amount of material is available.

Significant work has also been carried out in the mathematics of acoustic modelling of speech itself. Currently, the dominant framework for carrying out speech recognition is that of Hidden Markov Models (HMMs). These provide reliable results, and are used in most of the leading speech recognition systems. However, the mathematical assumptions made in setting up the models make several gross (incorrect) assumptions about the nature of human speech. Some research programmes aim to enhance the algorithms behind HMMs and add extra layers to the models, resulting in a mathematical framework which more correctly models speech. Other researchers are active in exploring alternatives to HMMs which do not suffer from their limitations.

Another current area of research is that of multimodality in speech recognition. This is partially an acknowledgement of the fact that face-to-face human communication is rarely carried out through speech alone. Gesture, gaze and gait, among other modes, are also important factors. Detection of lip movements can also aid in recognition: research in audiovisual recognition has been carried out for a number of years, and is yielding promising results.

F3. Computer-aided translation software – an overview and comparison of current packages

Overview

A CAT (computer-aided translation) tool is a technology that automates and assists the translation process for translators. According to some definitions this includes basic language tools such as spelling and grammar checkers, as well as translation memory (TM) software. Increasingly, controlled machine translation applications and other advanced utilities are seen as CAT tools for translator use.

Basic language tools

Basic language tools are generally included in office tools such as word processors, and are not marketed specifically for translator use. They are also language specific, and their usefulness to translators therefore depends on their availability in the languages they work with. All major and many minor languages are well-catered for with basic language tools, however, translators working with some less-resourced languages still have difficulty finding basic software tools for those languages. This seems less dependant on the size of the language community than the presence of computing and internet access for that community. Thus minority European languages (e.g. Welsh, Basque, Catalan) are well represented, but even languages spoken by large communities in Africa and the Indian subcontinent are under-represented in this area.

Translation Memory tools

Translation memory software on the other hand is not language dependent. TM applications work by storing both original text (in the source language) and their translations (in the target language). The software identifies previously translated segments, and displays them to the translator for re-use. These may be either full, 100% matches, or partial 'fuzzy' matches, where the text is similar, but not identical to that previously translated. The degree of 'fuzziness' may often be set by the translator, and a match of around 70% is often accounted to be the greatest help in speeding up the translation process, and ensuring accuracy and consistency. Old documents and their translation (called 'legacy translations') may be input into such systems, or a translator may build a memory from new. Memories may be shared by a team of translators, thus making the translations of one translator available to others. This is deemed to be especially valuable as translators move posts or retire, as their work is still available to younger, perhaps more inexperienced members of staff.

TM is most efficient when translating repetitive text, e.g. updating handbooks and manuals. However, it is also valuable in quality control, ensuring that all text is translated, and that there is consistency in the use of terminology and house style. TMs may be linked to dictionary resources, in the form of external bilingual terminology lists, and/or glossaries built up in-house.

Summary of TM basic functionalities:

- Presenting sentences to the translator in a convenient way, this is done by presenting a source segment and its translation as a unit.

- Storing of the translation units in a translation memory (TM) and automatically looking up the TM when a new segment has to be translated. Any result of the TM search is presented in a convenient way so that it can be re-used by the translator.
- Automatic look-up in terminology databases, and the automatic display and insertion of the search results.

Advanced CAT tools

Commercial TM software is becoming increasingly sophisticated. Market leaders are incorporating other features into the software to provide one-stop solutions to translators, especially at the top end of the market where large agencies need to manage a great number of translators and translation projects. Thus tools for managing workflow and documents, calculating word count and billing are increasingly integrated into TM products.

Other areas of development include catering for languages other than those written with a Latin script, and catering for translation between two writing systems, with the attendant challenges e.g. of a system that writes right to left, as opposed to left to right. Machine translation in a controlled environment is also being addressed, in the form of preparing a preliminary translation that will then be post-edited by a human translator. Such developments are however based on specific language pairs, as opposed to simpler TM which is language independent.

Comparing CAT tools

There are a wide range of CAT tools on the market. They vary significantly with price, functionality and supports for various file formats[30]. For example, WordFisher[31] is simply an add-on tool to Microsoft Word which is available for free but with very limited functionality. At the other end of the scale, SDL Trados[32] is a standalone application with a significantly greater number of features, not only assisting the translator in translating the text, but also offering other facilities such as cost estimation.

Some of the established players are better known than others. In particular, Déjà Vu[33], SDL Trados and Wordfast[34] are favoured by large organisations or institution. Other solutions, such as OmegaT[35], are emerging on to the market. OmegaT in particular is gaining popularity with freelance translators because it is available for free as an open source solution and is platform independent.

Due to the potential commercial impact of any pronouncement of the relative merits of the various CAT tools on offer, caution needs to be exercised in recommending any one solution over another. However, this is a question of much interest to translators, especially freelancers who are in charge of their own software purchases.

---

[30] There is a useful table comparing different CAT tools at http://en.wikipedia.org/wiki/Computer-assisted_translation.
[31] http://www.wordfisher.com/
[32] http://www.trados.com/en/
[33] http://www.atril.com/
[34] http://www.wordfast.net/
[35] http://www.omegat.org/

Training and support in the use of CAT tools

Training and technical support in the use of TM and other CAT utilities is an issue for many translators. Many TM vendors hold training seminars on the use of their own software, but there is a lack of provision at a more general level. Translators in Wales feel they have inadequate training in the technology, and technical support is sparse, both within organisations who have their own technical support staff, and even more so for freelancers and very small companies who do not have dedicated IT personnel. These issues are shown in the interview with Tegau Andrews (Appendix C3), and Focus Group discussion (Appendix E). It seems therefore that although CAT technology is fairly ripe in the international field, there is ample scope within Wales to develop training and IT support for CAT tools in the translation sector.

F4. Review of international best practice in machine translation

Overview

By reviewing contemporary literature on Machine Translation (MT), current best practice was identified, and a system conforming to these practices was evaluated, with the requirements of SMEs in Wales in mind. We found that, given a suitable bilingual text corpus, very effective machine translation can be achieved. Further research which could be of immediate practical benefit was identified, and relevant issues noted accordingly.

Machine Translation Paradigms

There are three major classes of machine translation.

Statistical Machine Translation (SMT): a bilingual text corpus is analysed to produce a statistical model of the mapping from a source language to a target language. Subsequently, given text in the source language, a most likely equivalent in the target language is found according to this model. SMT is currently the paradigm in which the majority of MT research and development is being undertaken.

Rule-based Machine Translation: a system of lexical, grammatical and reordering rules is created for a source-language/target-language pair. The rules are then applied to subsequent source text to produce translated output.

Example-based Machine Translation (EBMT): a bilingual text corpus is used directly for comparison against source text, and case-based reasoning is applied to create a translation.

Purported EBMT systems are often in fact hybrids of the above approaches to some extent, and systems exist using various combinations of these techniques and others.

Evaluation of an SMT system

One of the current leading SMT systems is Moses, a factored phrase-based beam-search decoder, a free, open-source project licensed under the LGPL[36]. An online demonstration of Moses is available[37], using one of the most popular and comprehensive freely available corpora, the European Parliament Proceedings Parallel Corpus (Europarl)[38].

Moses was evaluated using a bilingual English-Welsh text corpus. Other language pairs would also be useful to SMEs in Wales, in dealing with minority language communities as well as international communication. But since SMT uses a bilingual text corpus, and not specific rules about the languages in question, most of the issues would be similar in all these cases.

[36] http://www.statmt.org/
[37] http://demo.statmt.org/webtrans/
[38] http://www.statmt.org/europarl/

Unfortunately the Europarl corpus does not include the Welsh language because Welsh is not an official working language of the EU. The availability of an appropriate corpus is of primary importance for SMT, and in our opinion, likely to be the limiting factor in the development of effective MT systems, particularly for lesser-resourced languages.

Therefore, for evaluation purposes, a corpus was assembled from the Welsh Statutory Instruments, which are legislation published in English and Welsh in accordance with the Government of Wales Act 1998[39]; this makes a useful test system because legislative language is highly regular and specialised. The more comprehensive Welsh Assembly Proceedings corpus assembled by David Talbot contains actual spoken language and as such would likely produce better results in a variety of more general contexts. The Cronfa Electronig o Gymraeg[40] assembled by the Language Technologies Unit at Bangor University is not appropriate for this purpose because it is monolingual (but see below).

Moses treats each sentence independently, so sentence boundaries were identified and the English and Welsh versions matched up, using a custom script written in the Python programming language. Note that this sentence-by-sentence approach implies that a corpus whose translations are relatively non-literal (with sentences being combined or split) will be drastically less effective.

The statistical models were generated from the corpus by a computer program running overnight. Note that more time would be required for a larger corpus, and that this could conceivably have implications for iterative development, where the software developer repeatedly makes small changes to the system and then tests it. If the developer must wait several hours to try out each small modification, the overall project could take a long time; therefore, one of the software development techniques which avoid this problem would have to be used.

Analysis of Translation Quality

The quality of translation obtained is highly dependent on the bilingual corpus used -- in this case, legislation, which is written in specialised formal language. The system performs well when translating text of this type. Here is a hypothetical sentence from a non-existent Regulation, using many common phrases:

> The Welsh Ministers make the following Regulations in exercise of the powers conferred on the Secretary of State by section 1 of the Local Government Act 1972[1] and now exercisable by the National Assembly for Wales.

And here is the suggested translation:

> Mae Cynulliad Cenedlaethol Cymru yn gwneud y Rheoliadau canlynol drwy arfer y pwerau a roddwyd i'r Ysgrifennydd Gwladol gan adran 1 o Ddeddf Llywodraeth Leol 1972[1] ac sy'n arferadwy bellach gan Gynulliad Cenedlaethol Cymru.

[39] http://www.opsi.gov.uk/legislation/wales/w-stat.htm
[40] http://www.bangor.ac.uk/~cbs204/ceg/newidiadau_i_dagiau_ceg.html

In this case the translation is perfect. In other cases, there are errors but the suggestion is a useful starting point and could serve as an effective enhancement to a human translator, significantly increasing the rate at which translation work can be performed.

At the opposite extreme, the system has almost no idea how to translate the following sentence:

> Sheep only eat grass

The output is:

> Dafad eat grass yn unig

The words “eat” and “grass” actually occur repeatedly *in English* in the “Welsh-language” legislation, because it quotes monolingual legislation from the UK Parliament which is being amended[41]. Text of this type should be omitted from the corpus to avoid this happening.

We conclude that, given a suitable corpus, SMT can generate good-quality English-Welsh translations and could be very useful to a human translator.

Areas for Further Research

It is clear that efficient SMT requires a large corpus of high-quality bilingual text. The system is most effective when used on material similar to that in the corpus. It may be useful to have several corpora for different subject areas. There is a great deal of bilingual text available from public institutions in Wales, which could be used given appropriate resources to collate and prepare the material in the appropriate format. Both manual collation and technological solutions such as automatic alignment (matching up sentences) would be of benefit. **Assembling an appropriate bilingual corpus is the major factor in achieving effective English-Welsh machine translation.**

English-Welsh MT tools would be more useful if integrated into translation memory frameworks, a number of which are in widespread use in Wales, both proprietary and free/open-source[42]. When a document is being translated, the machine translation can then conveniently act as a starting point for further refinement by a human translator. This is a step beyond current practice in Wales, whereby translated texts are stored and suggested if they occur again; MT can make new suggestions for phrases which it has never seen before.

Translation involving other languages would also be useful to SMEs in Wales, e.g. Polish, Chinese. In many instances the typical usage might be somewhat different; for instance an imperfect machine translation might actually be of direct use to a person

[41] See, for instance, http://www.opsi.gov.uk/legislation/wales/wsi2005/20051156w.htm#7

[42] e.g. Trados (http://www.trados.com), Wordfast (http://www.wordfast.net), OmegaT (http://www.OmegaT.org)

who understands little English, though caution should be applied not to create unrealistic expectations[43].

Effective tools[44] already exist for identifying Welsh parts of speech (and, for instance, recognising that “cael” (get) and “cafodd” (got) are different forms of the same word), and output from these could enhance Moses's word-recognition capabilities. The monolingual CEG corpus could also be used to improve word recognition.

Due to time constraints, many of Moses's default “models” were used. On example is that translations are preferred where the word order of the translated sentence is close to that of the original text. We suspect that there may be better models for Welsh-English translation, because of significant differences in word order (VSO/SVO, pre/post-positional adjectives, periphrastic verb forms etc).

[43] A number of very public Welsh mis-translations (e.g. See http://news.bbc.co.uk/1/hi/wales/5341646.stm) have resulted from a misleadingly marketed program which performs rudimentary and erroneous word-by-word translation (http://www.tranexp.com/#InteractiveTrananchor)

[44] The lemmatizer component of Cysill, developed by the Language Technologies Unit, Canolfan Bedwyr, Bangor University

Appendix G. Accounts of visits to internationally regarded laboratories in SALT fields

G1: Interview with Professor Martin Russell, University of Birmingham, 28th January 2008

The interview was conducted in Professor Russell's office in the Department of Electronic, Electrical & Computer Engineering (EECE) at the University of Birmingham.

Introduction

Professor Martin Russell is the Head of Department at EECE. He came to the University of Birmingham in the late 1990s, and was previously a researcher in the Speech Research Group at the Defence Research Agency (later 20/20 Speech) in Malvern.

Current projects

Current projects being undertaken by Professor Russell include:

- Intermediate layer models (EPSRC funded). This is a project to enhance the way in which mathematical models use knowledge of the constraints of human speech. These models are used in synthesis and recognition. By introducing an intermediate layer to these models, in between the lower-level mathematical state layer and the higher-level speech layer, it is anticipated that the performance of speech recognisers will be improved. Specifically, the use of intermediate layer models is likely to reduce the time taken to adapt existing speech recognisers, on-the-fly, to recognise a new person.
- Noisy speech enhancement. Recognition of noisy speech has always been a challenge to speech recognition systems. This problem is of growing importance as such systems find themselves applied in new and diverse environments, such as automotive and roaming locations (e.g. mobile phones). Professor Russell had PhD-level projects researching various ways of enhancing speech to improve the performance of speech recognition in such situations.
- Accent recognition. The Accents of the British Isles corpus has been collected as part of investigations to establish how speech recognisers perform with variously accented British speech. More information on the original corpus can be found here: http://www.thespeechark.com/abi-1-page.html . It is currently marketed by The Speech Ark, a company spun out of the university's research work. Funding for database collection has been gathered from a variety of sources, including most recently, Advantage West Midlands for a collection of spontaneous telephone conversations carried out by individuals around Birmingham.
- Multimodal projects are particularly important to the work carried out in Birmingham. This is partly due to their innovative nature being favourably viewed by research councils. Multimodal projects, however, are also argued by Professor Russell to be a more intuitive way of conducting research in speech technology. In our conversation, he argued that dictating to a computer is a highly artificial form of communication. It does not reflect our normal everyday communication with other human beings, where we use gestures, movements and other non-verbal modes to enhance the understanding of the speech being

uttered. Multimodal projects in Birmingham include one jointly with the Department of Psychology, which uses knowledge from eye-tracking devices to enhance speech recognition.

Previous collaborations

One collaboration pertinent to SALT Cymru was undertaken by Professor Russell in the late 1990s. This was a TCS (which would now be a KTP) between him and the Productive Play Company (now Awen) from Cardiff. The project looked at ways of recognising children's speech in order to incorporate speech recognition systems in children's teaching aids in classrooms. This project completed in 2000 and no work has followed on from this.

Future directions

In discussion about future directions in speech and language technology, Professor Russell was keen to emphasise the importance of multimodality. He saw speech being normalised in research, and becoming one of many modes of communication to be recognised in conjunction with each other. He postulated that the reason for the comparatively low take-up of desktop dictation software by end users was the artificiality of dictating into a microphone. He thought that researching speech as a mode of communication, enabling gesture recognition and similar techniques, might break down this artificiality and lead to greater user acceptance of the technology.

G2. Interview with Dr Lori Levin, Carnegie Mellon University, Pittsburgh USA, 27th December 2007

This report follows the visit of Delyth and Gruffudd Prys to the Language Technologies Institute, Carnegie Mellon University, in December 2007.

The Language Technologies Institute conducts research into many areas of SALT (see Appendix A). In addition to fundamental and theoretical research, the LTI focuses on large-scale challenges of consequence to industry, government and society at large. They see no conflict between engaging in highly speculative non-traditional research on the one hand (e.g. investigating the language of dolphins), and practical government or industry projects on the other (e.g. speech to speech translation on hand-held devices). They also work with minority languages, including indigenous languages of the Americas, and emphasise social goals as well as catering to the needs of business and industry.

The LTI has had links with the LTU at Bangor University since 2004. In 2006 DR Lori Levin from the LTI visited Bangor University and gave a talk on the machine translation research at the LTI. In 2007 one of her graduate students, Christian Mason, presented a paper at a colloquium organised by Bangor University's LTU at the 11th International Conference On Minority Languages (ICML 11) Pécs, Hungary, July 5-6, 2007 on Language Revitalisation through Multimedia Technology.

Industrial Programmes at the LTI
The LTI collaborates with many national and international industry partners such as IBM, Fujitsu, Siemens and Hewlett Packard. They are also able to offer affiliateships (where industry participates in the results of LTI research), in-residence fellowships (full immersion training in new technologies), customized intensive courses, and sponsored R&D projects for industry and government.

Academic programmes at the LTI
A Master of Language Technologies (MLT) is offered that cover linguistic and statistical approaches and basic computer science. The three language technology areas covered are Machine Translation, Information Retrieval, or Speech Technology, and the student usually specializes in one of them.

A research-based Ph.D. in Language and Information Technologies is also offered at the LTI. Students must also successfully complete at least eight courses while working towards it, giving them a broad grounding in areas such as Linguistic, Computer Science, Task Orientation, and Statistical/Learning and lab work in a choice of Speech, Machine Translation, Information Retrieval, and Natural Language Processing.

Graduates and former students from the LTI are much in demand by employers, supplying specialist skills to industry and government.

Attracting prospective students
In an effort to attract top quality students to enrol at the university, CMU takes part an annual competition for "high school students interested in Language, Math, Computers". This takes the form of a North American Linguistics Olympiad, where successful students go on to take part in an international science olympiad (see http://www.naclo.cs.cmu.edu/). The LTI representative, Dr Lori Levin, was very keen to

recommend this model to Wales and urged the LTU at Bangor University to organise such a competition in the schools of Wales.

Appendix H. Account of conference visit

Report on LangTech 2008, Rome, 28th-29th February 2008
Attended by Dewi Jones and Rhys Jones

1. Key presentations attended/posters viewed

Kimmo Rossi – Putting HLT (human language technology) research and technology into action for European multilingualism

- The European Union's Framework Programme 7 (FP7) does not yet fund language technology resources projects. The trend in FP7 is towards funding integration and mainstreaming of language understanding in interactive systems and knowledge management.
- In the meantime, the final outputs of the last calls from FP6 and other projects on language technology resources are being collected into the CLARIN and FlareNet projects. Results or progress of these two initiatives will feed into forming FP7 calls in 2009 and 2010.
- Diversity means business. And Europe is not a fortress. It needs to connect with China, India, Russia and Arabic countries.
- Recent advances in language technology already offer great potential to businesses, but a lot of work still needs to be done. Further efforts are needed to put the state-of-the-art language technologies into productive use.
- While the major languages are well equipped with language resources and tools, there are still gaps in the coverage of some less widely spoken languages. Efforts are still necessary to make languages more equal.
- Lots of projects/research is being conducted in multimedia settings.

Language Technology in Search – assorted presentations

A number of search technology providers such as Google, Microsoft and various start ups presented how LT is increasingly used to improve search functionality e.g.:

- Summarization – especially important for delivering content to mobile devices where time and space are at a premium
- Natural language processing (NLP) techniques can filter ratings - ratings that are offensive, off topic and/or spam. (i.e. automate human editors.)

Smaller companies, such as Empolis are able to employ language technologies for bespoke intelligent search solutions. A lot has been done to date but more work is needed, in particular

- Employ part of speech tagging, semantic annotation, entity extraction and cross lingual search technologies
- Better and smarter integration into applications.

Traditionally NLP has been regarded as a key component for understanding and forming natural language queries such as "I am interested in NLP used in information retrieval". However, NLP is increasingly needed more in understanding search results. Further

work on NLP in natural language querying is redundant and misguided since people have adapted to the key word searches interface provided by engines such as Google.

A presentation by Google informed us that :

- Google uses lexical semantics to deal with synonyms and synonym expansion.
- Machine learning (ML) is used instead of building up thesaurus dictionaries. ML learns what are the related terms.
    - e.g. If some searched for “used cars” and the user clicks on a page with “used automobiles” then the system can learn that the two are synonyms.
- Google’s machine learning approach is not perfect, but it is getting better and it might have the critical mass of users (i.e. in the order of millions) to statistically produce proper thesauri without the need for humans.
- Google provide some cross-lingual search capability between Russian and Arabic.
- Google will use more NLP and synonym data to give better search results.
    - e.g. Improve the diversity of search results. Rather than give the top 10 hits / page rankings, they would give results from the different senses of the search word/term. Google said this is needed and is being developed/invested in at Google.

Hakia is a relatively new search engine which uses semantic technologies. It develops and provides semantic search solutions and can answer questions written in plain English.

- Synonym expansion is key to Hakia’s operation – they have utilised 20 years of work in producing an English thesaurus.
- Hakia do not have the critical mass of Google's for utilizing machine learning in creating its dictionaries.
- Hakia commissioned subject specialists and semanticists to create the basic language technology resources for their synonym expansion.
- They have also built their own language models and ontologies for certain domains and types of language e.g an offensive language model.
- Initially semantic search engines like Hakia will be limited to certain specific domains. Google and Microsoft engines are general language and have critical mass for employing machine learning. (a domain may not have the scale to employ machine learning)
- Hakia have successfully used large public resources such as Wikipedia to test and validate their technologies.

Voice Search for Mobile Devices - Microsoft

Microsoft has recently launched (in the US) their Live Search for Windows Mobile service. This uses speech recognition to provide localised knowledge retrieval ‘on-the-go’. The service is targeted at the sort of information that users might want to retrieve in mobile locations, e.g. local maps, films and restaurant details.

To simplify the deployment of the technology, a simple 'push-to-talk' system is used, and the speech sent back to Microsoft's systems. Hence all recognition takes place on Microsoft's servers rather than the phone itself.

The importance placed to mobile devices by Microsoft reflects the ubiquity of the technology. It was projected that by 2010, there would be 250 Internet users and 200 PCs in use for every 1000 people worldwide. In comparison, there would be **500** cellphone users per 1000 people.

A challenge for the project is how to build a revenue stream for mobile search similar to the highly successful ones for desktop Internet search. The multimodal nature of mobile search (text and images on a screen, voice input) presents challenges but also opportunities in this area. For instance, advertising information can be sent to a user by SMS, making it more persistent. It was even suggested that demographic clues might be gathered from speakers (i.e. their speech analysed to attempt to determine their age and gender) and appropriate advertising targeted as a result.

One of the presentation's main conclusions was that speech recognition, in this context, enables applications that would not be practical otherwise – given the level of reliability claimed for speech recognition in the presentation, it provides a simpler interface than a mobile keypad for entering search information.

Challenges of Speech to Speech Translation in the context of Human-Human Communications – Alex Waibel

This presentation dealt with the major European/American project CHIL (Computers in the Human Interaction Loop). It develops and integrates a range of technologies, such as speech synthesis, speech recognition, machine translation and gaze detection, to aid communication in meetings where participants do not share a common language.

It is a sign of the maturity of many areas of SALT that the integration of e.g. speech recognition and machine translation is now considered a viable area of research. Quoted in the presentation was that the word error rate for speech recognition, in a meeting environment, has dropped from 30% in 2004 to 7% in 2007. This comparatively low error rate for speech recognition makes its integration with machine translation more feasible, as the translation stage is more likely to receive high quality text as an input, thus having a positive knock-on effect on the quality of the eventual output.

The size of the potential market, savings and opportunities was underlined by the statistic that €1.3 billion has been spent on translating government meetings in the European Union. Any technology enabling this to be achieved more quickly and effectively would have a highly positive effect.

Understanding the Market Movements in Network Speech

- Network speech was defined as the use of speech technologies, both speech recognition and text to speech in public facing automated services.
- The global and American market movements were presented.
- The biggest trends are to reduce costs and to increase scales of services**.**

- As part of long term strategies for automation and customer care strategies, companies are investing in automation rather than in offshoring call centres
- Use of software as a service or on demand is gaining ground in servicing the expansion in investment.
- Investments will increase in the next 5-10 years – current basic application of network speech will be improved and optimized through agile methods.
- Market sizing projections – spending on speech recognition to be $3.2bn in 2010. In 2005 it was $1.2bn. This includes costs such as hosting, related services, applications (in fact these form the majority of the costs/revenue). The core technologies for network speech will constitute/count for only 30% of the projected revenue. Companies using speech recognition will make more money than those developing speech recognition.
- Companies make more money on associated services and hosting.
- Adoption and uptake will continue across all markets in the US through the next five years. When compared to the classic adoption curve, speech has crossed the chasm – from tourism, financial services to Health Care, retail etc.
- Network speech has transitioned from a cool technology to an applied business solution. Open standards such as VoiceXML have driven the market. Speech is becoming an accepted UI and companies are establishing long term positions using the technologies.
- UK accounts for a limited amount of worldwide deployments of network speech – 11% compared to 56% in the US.
- Supported languages are: British English 9%. (so 2% in the UK use the American English speech technologies). American English accounts for 46% of languages supported and Spanish (Americas) 27%
- Companies have benchmarked success and will continue to optimize application performance.

Venture Capital and Language Technologies.

- Two ingredients regularly coincide: a new technology and an old human need. Any technology that met basic human needs would be embraced: people need to communicate; people are thirsty for information and entertainment.
- The existing market tells us little about what would happen if translations were available in real time and at reasonable cost. For the whole of Europe it is a strategic necessity to have human language technologies available that facilitate cross lingual communication and information exchange to the greatest extent possible. LT is an economic, political and cultural necessity.
- The cost of localization into a new language does not correlate with the number of speakers but is more or less fixed. There are as a consequence primary languages, the ones in which one must have a basic offering, and then there are secondary and tertiary languages of minor commercial relevance. Market forces penalize those languages with small speaker populations or with associated weak economies.
- Any technology that reduces this cost supports our cultural heritage.
- Language barriers place a burden on our economy as they add transaction cost to any activity that crosses a language border. The cost is small usually compared to the product cost (0.25%-2%). Cost is often indirect in the sense that it inhibits the creation of value.

- In 2005, the language industry generated $8.8bn in revenue. Two important segments that are driving growth are
    - multilingual websites and
    - software localization.
- More than 5000 companies worldwide with > 5 staff contribute to this market.
- 4 out of the top 20 companies have worldwide HQs in the UK (in and around London and in the North of England). They have $289m in revenue and 2072 employees.
- In Ireland there are two companies with worldwide HQs (though others have offices in Dublin also) with $36.2m in revenues and 283 staff.
- Satellite offices from all or some of the other top 14 companies count for more staff and economic activities in these areas of the UK and Dublin.
- Technologies currently exploited by these companies include – translation memories, terminology databases, multilingual CMS and tools for software localization.
- Putting aside quality issues, two major market drivers now exist for machine translation (MT):
    - Cost - driving costs down by a significant order of magnitude will boost the use of translation as the latent demand as yet is far from satisfied.
    - Accessibility - accessibility means the latency of obtaining a translation. Real-time exists but is costly.
- MT will remain inferior to human translation for many years. The market will be dominated by both. Human translation where quality is an absolute necessity (or where machine translation is not viable). MT will dominate at the low end of the market and new and in new emerging markets which will emerge as a consequence of the availability of low cost translation technology. MT will move up market eventually as quality and performance improves.
- The economic conditions for LT are different between the EU and the US.
    - In US, it is speech recognition and dialogue technologies in the area of customer self service that are growing. i.e. the automation of human call centres. The US has had tedious DTMF and rigid menus for a while and the shift to more speech enabled services is seen as an improvement in the companies' service offerings. In the EU, call centre services have been available for a number of years and have provided a high level of service. The introduction of automation would be seen as a regression. There are also better economies of scale in the US for speech recognition since only one or two languages (i.e. Spanish) is required.
    - The EU offers converse economic conditions that favour machine translation (and speech to speech translation) since many languages are spoken. In Europe, doing business means being multilingual. The EU is in urgent need and at the same time is in a privileged position. Given the technology, it offers an option to commercialize MT into other world regions. It cannot expect its needs to be satisfied by other these other regions.

2. Contacts made

Tigran Spaan, GridLine BV

Gridline is an IT consultancy company based in the Netherlands, which in addition offers Dutch language services and technologies. The company draws on work done on Dutch terminology for providing solutions on content management, document management and knowledge retrieval. Their products contain basic language resources for Dutch such as lemmatizer and part-of-speech tagging components

Their Dutch language service and product offerings put GridLine in a very competitive position for securing large public sector clients in the Netherlands such as the Dutch government. Products such as their GridWalker Thesaurus Tools have been keen in winning some major clients and getting a foot in the door for further IT consultancy.

Recently they have been commissioned by the Dutch Language Union to commence a project to look at 'Clear Language in Government', which aims to make government communication more understandable to the general public by using clear, accessible language.

GridLine has good relationships with various Universities in the Netherlands that are conduct research into Language Technologies. They are helping GridLine develop its next product for mood analysis. i.e. to analyse texts on the web e.g. blog entries, to determine what people are saying about you and/or your product/brand. The solution will be developed for English and Dutch.

3. Findings for SALT Cymru

The LangTech conference was a successful conference partly hosted and sponsored by ForumTAL [45] . ForumTAL itself is a permanent forum set up by the Italian government for promoting research and development of highly innovative language technologies amongst Universities and SMEs in Italy. It also helps promote Italian language technology through the EU and worldwide.

LT is at a level of maturity for supporting sufficiently the EU and 'primary' languages such as English, Spanish, German, Italian and French. The LT are being improved and optimized with the scope being expanded into an increasing number of applications. The developments in multilingual LT are allowing for support for a growing number of lesser resourced languages being viable and sustainable.

It was hinted that the next phase or evolution of language technologies will push further multilingual capabilities and generics. Further pushes into automation and into limiting costs to affordable levels will come to the widest scope of languages used by individuals and organisations worldwide.

These lesser resourced languages provide a long tail of languages. When looking at the top 30 languages in the world, the numbers of speakers of lesser resourced languages outnumber the number of speakers of all the primary languages[46] . With a longer tail the difference will be even bigger. Some of these languages represent emerging economies.

[45] See http://www.forumtal.it – set up as a permanent forum for promotion research and development of highly innovative language technologies in Italy and of Italian LT worldwide.
[46] See http://unicode.org/notes/tn13/ GDP by Language

Lesser resourced languages represent a fertile area of opportunities for research and exploiting language technologies. However, as has been seen with research and development of language technologies for Welsh, LT support for lesser resourced languages present unique challenges, not only in scaling up multilingual technology but also in cross disciplinary aspects related to adoption of language technologies.

- LT for Welsh has a more daunting challenge than LT for larger languages. It must operate and be effective to higher levels of accuracy and cost less to resource and support than other languages to date.
- LT support for a lesser resourced and spoken language has to be of excellent quality. This is not so as to satisfy purists but because the language is less robust to inaccuracies and 'noise'. If not robust enough, 'noise' caused through inadequate language technology begins to stick and permeate through the language and is not part of the human evolution of a language, but rather an artificial side-effect

In the next evolution of language technologies each lesser resourced language has a contribution. Welsh has a significant contribution from its experience and provision that is invaluable in informing future developments. Wales and Welsh LT is in a privileged position.

Appendix I. An evaluation of relevant open-source software and standards with a view to enhancing them for use in Wales in a pre-competitive research stage and deploying them for take-up and further development by industry in a non state-aid environment.

I1. Festival

Relevance to the field
In the SALT Cymru survey, it is noted that speech synthesis (text-to-speech) is one of the key areas both for users and developers. In particular, *speech enabled communication aids for disabled users and those with specific needs* was the SALT category developed by the greatest proportion of SALT developers participating in the survey. Further, 18% of developers said they would be interested in developing speech synthesis further in future.

Festival provides a useful system for the development and deployment of speech synthesis. It is available as open-source software with very few restrictions, and its comparatively liberal license means that it can be used commercially without the need to expose industrially valuable source code. Further, Festival is a modular system, and basic modules exist within its framework to enable most of the required functions of a speech synthesis system. This means that development time can be targeted to improve and enhance the features required by the developers, without necessarily needing to develop a complete system.

Ease of use
Festival, in its native form, runs from the command line, thus knowledge of its commands is required in order to be able to use it. An ordinary user should not be expected to have this level of expertise with the system in order to be able to use it, and thus a simpler interface is sought.

Such an interface exists, and has been developed by Bangor University's Language Technologies Unit. It works with Microsoft Windows through the operating system's Application Programming Interface, and integrates Festival with Windows so that it can be used with any speech-enabled application that conforms to Windows' standards. This allows a variety of screen readers and similar assistive technologies to work with any Festival voice. The interface can be downloaded freely under the same open-source license as Festival itself.

Ease of development
As a modular system, Festival can be used with little or no extra development. Festival voices exist for English, Welsh and many other languages, and can be deployed by simply downloading the required voice modules and executing the relevant voice commands. The voices can also be operated in a more user-friendly manner through the Windows interface, as explained in the previous paragraph.

Festival contains standard modules that provide the basic functionalities of a speech synthesis system, including speech output, phrasing and intonation. It could, therefore, be deployed with little or no additional development beyond its integration with the developers' speech input and output systems.

The main challenge to Festival developers is improvement of the voice quality from that which is readily available. Most of the voices currently available for Festival (including all those available in Welsh) are diphone voices. These are suitable for screen readers and other applications where the user is likely to listen to the voice for extended periods of time and hence familiarise themselves with the voice's qualities. Diphone voices are, however, less suitable for telephony applications, and other scenarios where the user will only have limited exposure to the voice, and where utterances are less likely to be repeated. For these situations, unit selection voices are the preferred synthesis technique.

A framework to develop unit selection (and other) voices within Festival already exists. Called Festvox, it uses a series of scripts and pre-written code to reduce the length of time, and the amount of programming expertise, required for voice development. Literature has been written on developing unit selection voices within this framework for languages that did not previously possess such a resource[47], and as such, the process for developing a new unit selection voice is reasonably well-trodden and well-described. It does, however, require a developer with a level of expertise in speech synthesis (or time to be set aside to learn the fundamentals of the field) and would need, at the very least, six months of one person's development time, plus additional time and a speaker to record the required unit selection speech database.

In developing voices for Festival, it is worth bearing in mind that it was principally designed as a *development* framework rather than a system for deploying voices. It is true that the computational load of a Festival voice is not excessive, and Festival voices can be run on a standard PC with relatively modest processing power. However, if a voice is to be run on a low-power embedded or hand-held system, alternatives to Festival should be sought. In particular, Flite, developed by many of the Festival team, offers a speech synthesis system with a smaller footprint, suitable for a wider range of devices. A mechanism exists to transfer voices from Festival to Flite for their deployment.

Potential worth to Welsh SMEs

Festival offers a ready answer to the problem of speech synthesis for those requiring to develop it. Acceptable solutions are already available for English-language and Welsh-language synthesis using the package. These solutions are not universally applicable, but do however give results that are suitable for the majority of PC users that are likely to rely on speech synthesis for their day-to-day computer use.

It is evident that those SMEs willing to invest time and effort in Festival development should reap rewards from doing so. To a large extent, off-the-shelf solutions are available for most of the likely deployments within Wales. Development of higher-quality voices for Festival, while challenging, should not be an insurmountable problem, especially if time is invested within companies in building knowledge of speech synthesis techniques. It is felt that a significant competitive edge would be gained by a company able to develop, for example, a high-quality telephony system using a newly developed Festival unit selection voice.

Festival is also well-integrated with Windows, meaning that most of the research time for new Festival developments can be taken up in the core work of voice building, rather

[47] For the Amharic language, see, e.g. www.cs.cmu.edu/~awb/papers/ssw5/amharic.pdf

than being occupied in integration work which is not central to the development process. This has the effect of reducing a company's time to market on Festival development, and increasing profitability.

I2. Sphinx

Relevance to the field

Speech recognition has been found to be a key area for the SALT Cymru survey. It was the second most popular area of SALT in terms of current development, and was considered very important or fairly important by 47% of developers. Additionally, 41% of developers stated that they might develop speech recognition in the future.

In searching for SALT that might aid Welsh SMEs, which are nevertheless able to be developed within a non-state aid environment, a useful model is that of free software development. Such software is able to be developed by academic institutions via grant or other funding, and collaboration in development is facilitated by access to the underlying source code. While the source code to any software can be considered an industrially valuable asset, certain open-source licenses (termed 'BSD-like') allow further commercial exploitation without exposing this commercially valuable source code. Hence, such licenses provide an ideal platform for developing in an environment that favours no one company or organisation, but nevertheless retains the ability for any interested parties to commercially develop any results that arise.

In developing speech recognition, it is not considered reasonable that a significant amount of work should be expended in coding basic algorithms that are well-known to the research community. It is recommended that use should be made of pre-existing packages, in which the algorithms are already coded, so that time can be spent developing applications within the existing frameworks.

The main packages for speech recognition, available at no cost, are **HTK** (developed originally by Cambridge University's Engineering Department) and **Sphinx** (a series of training and decoding decoders developed by Carnegie Mellon University). Of these, only Sphinx can be freely redistributed with end applications. It is free for commercial and non-commercial use, and has a BSD-like license. HTK would require the end-user to download and compile a version of HTK for their platforms.

In the majority of cases, it is not felt reasonable to ask an end-user to compile their own software to enable speech recognition. By this criterion, Sphinx should be the choice of recognition software.

Developing speech recognition for Wales

In developing speech recognition for applications in Wales, it is assumed that support is required for both the English and Welsh languages.

In English, speech recognition may be considered to be reasonably mature, as it has been the subject of development work for over 40 years. Of particular relevance to Sphinx is that acoustic and language models for English, trained on large amounts of data, are available for free download from an associated website[48]. It is stated that this 'may just work' for individual needs, though it is implicit that additional data and development would be required for specialised applications.

[48] See http://www.speech.cs.cmu.edu/sphinx/models/

For Welsh, the situation is entirely different. Very little work has been done on speech recognition in the language, which means that development of any practical speech recognition system requires, essentially, its development from scratch.

Any speech recognition development involves the following stages:

1. Definition of a recognition task.
2. Selection and preparation of training and test data sets.
3. Training an acoustic model on the training data.
4. For all but the simplest recognition tasks, training a language model.
5. Running the decoder on the test data, and deriving a recognition score for the task.
6. Optionally, modifying the acoustic and language models to improve the recognition result on the given test data.

The description above details an off-line speech recognition system, i.e. one which operates on files of data rather than recognising live speech. If live speech recognition is required, the following must also be addressed:

7. Running the recognition stage on live speech input.
8. Optionally, adapting the acoustic and language models given the speech input data.

Definition of a recognition task

The recognition task comprises the range of speech the recogniser is expected to be able to deal with, in terms of the speaking environment(s), the range of speakers and the complexity of language. Recognition results are likely to be higher for a smaller range of speakers and a more restricted vocabulary within the application.

The recognition task must be carefully defined before further development takes place.

Selection and preparation of training and test data sets

In any speech recognition development, there is a trade-off between the amount of data used in training a recogniser and the quality of the eventual recognition result. A simple recogniser that distinguishes between a dozen or so words may only require a few minutes of training speech. At the other extreme of complexity, in most broadcast news recognition tasks over 100 hours of training speech are typically required to achieve about 30% word error rate in recognition.

In order to train acoustic models, the speech data must be segmented at the phonetic level. An automatic process of forced alignment will be used for this, involving individual utterances being aligned with a phonestring derived from their word-level transcriptions. Therefore, a pre-requisite for the forced alignment process is a phonetic transcription of each word in the training data. This will be derived from existing lexica developed for speech synthesis.

Forced alignment can begin once a transcription at the phoneme level is available for each word in the utterances. These transcriptions can be derived using freely available

phonetic dictionaries (lexicons) for Welsh, supplemented by the use of freely available letter-to-sound rules for those words not present in the lexicons.

Forced alignment works by taking an initial estimate of the segmentation of the utterances given their phonetic transcription. The initial estimate assumes that all phones in the utterances are of identical duration. Phone models are trained on this initial uniform segmentation, and then used to derive a new set of time-aligned transcriptions, which are found to be a slightly better match to the actual alignment than are the initial uniform segmentations. These new transcriptions are used to derive a further set of transcriptions, which are again found to be a closer match to the actual alignment. This process of convergence is repeated over a number of cycles to produce an accurate, automatically derived, time-aligned transcription of each utterance. Manual checking of a proportion of the results is essential, in order to determine

Training an acoustic model on the training data

Typically in speech recognition, acoustic modelling takes place at the phoneme level (i.e. at the level of the individual sounds of words). Individual Hidden Markov Models (HMMs) are used to model each phone in the language to be recognised. Standard techniques exist to train these models, and these techniques are implemented in Sphinx.

The only decision to be made at this stage of development involves the definition of a set of acoustic models for Welsh, i.e. the definition of the individual speech sounds which are to be modelled. Earlier work by the Language Technologies Unit at Bangor University has defined a set for the North Welsh accent[49]. This however includes a large number of sounds. Depending on the amount of training data available to model each sound, it may be found advisable to use a simpler set of sounds, such as those used in the South Welsh accent. Any sounds present in the North Welsh accent would then be mapped to their equivalent in the South Welsh accent.

Training a language model

In speech recognition, a language model is used to reduce the number of possible candidates for the output text. It typically determines the probability of words in given contexts – usually, of a given word being followed by another given word or by a longer sequence of words.

A language model is trained from a large corpus of text suitable for the task in question. In the case of Welsh, two possibilities currently exist for this:

- CEG (approx. 1 million words)
- Crúbadán (approx. 95 million words)

While it would appear that the larger size of the Crúbadán corpus would result in a language model of higher quality, this is not the only criterion. The CEG corpus has been derived from primarily literary and newspaper texts. They have been quality-checked to a standard believed acceptable for most uses. Despite Crúbadán's larger size, its use may not result in a better recognition result. It has been collected from a broad selection

[49] See http://bedwyr-redhat.bangor.ac.uk/svn/repos/WISPR/Documentation/Technical/phoneset-wispr-welsh.pdf

of web text. As such it will probably contain many errors and text from other languages, notably English.

In the context of a language model, the number of errors in Crúbadán is less important than the type of those errors. Non-systematic errors in Crúbadán (e.g. specific words being largely spelt correctly, but occasionally suffering from typos) should not significantly affect the quality of the language model. Any n-gram based model would give a low probability to these errors, so they should not significantly influence the system's output. On the other hand, systematic errors in the corpus (e.g. a specific word being followed by another specific word, but incorrectly mutated) could affect the language model. If words occur regularly in specific incorrect contexts, the probability of those n-grams in the language model will consequently be higher, and the output of the system could be adversely affected.

In purely practical terms, the availability of CEG makes it more suitable for the rapid development of a language model in the first instance. However, it is strongly recommended that Crúbadán be investigated in future developments. It is suggested that after using a CEG language model within the speech recognition development, a Crúbadán language model should be substituted, and the results investigated.

Running the decoder on the test data, and deriving a recognition score for the task

At this stage in development, the acoustic and language models have been fully trained. A decoder can thus be selected, which will be given the trained models and input speech, and which will attempt to transcribe the text of the input speech.

Several versions of decoders have been developed as part of the Sphinx projects. They vary in methodology, in the programming language used to develop them, and in their levels of maturity. The most recently developed decoder is Sphinx-4, which is written in Java. However, it is aimed at off-line processing rather than live applications. Sphinx-3 is an older decoder, written in C and originally aimed at off-line processing, but which has recently been further developed for live recognition. It is regarded as the most accurate decoder developed as part of the Sphinx project.

The output of the decoding stage will be a set of transcriptions. To achieve a recognition score from these, a separate program is run. Sphinx includes a program that derives recognition scores which comply with the NIST standard, allowing results to be compared with speech recognition developments in other languages.

The decision of what constitutes an 'acceptable' recognition score is open to question. The state of the art in recognition of broadcast news bulletins is a word error rate of about 25% (a word recognition score of 75%). Such systems have normally been trained with 100 hours or more of speech data. However, any initial system for Welsh will have been trained on a much smaller amount of training data, and this should be borne in mind when comparing results.

Modifying the acoustic and language models to improve recognition of the test data

Modifications to acoustic modelling

A significant part of speech recognition development has involved adjusting various parameters within the recognition models and investigating their effect on the recognition result. The parameters to be used have by now been well-researched, and most speech recognition systems reported in research literature now use HMMs with three states per phone and MFCCs of about 13 orders with their first and second differential. However, some slight improvements in recognition results may be achieved by altering the topology of the HMMs – in other words, which transitions are allowed between their three states.

Varying the complexity of the acoustic models may also be necessary. A more complex acoustic model (technically, a greater number of Gaussian components per state of the HMM) allows a greater amount of training data to be modelled with increased accuracy. However, if there is not enough training data, then lessening the complexity of the acoustic models usually improves the recognition score.

Provided sufficient training data is available, the modification of the phone models to triphone models may improve recognition accuracy. Triphone models take into account the preceding and following contexts of the individual speech sound (the phone) being modelled. This results in an increased number of models and hence an increase in the amount of data required to train them. However, techniques exist to 'tie' the states of triphones with similar contexts, allowing the training data to be shared between two or more models. This reduces the total amount of speech data required to adequately train the models.

Modifications to language modelling

Analysis of the word patterns found within the text output may reveal errors and inconsistencies in the language model used. In particular, one unexplored area for Welsh is that of mutations and inflections, and how the text output would reflect those. It may be necessary to include mutations as alternative pronunciations of the same word form within the lexicon.

Other modifications

Additionally, the decoding stage may be made more accurate by increasing the word insertion penalty. In speech recognition, it is sometimes found that while all the words in the speech input are present in the text output, additional words have been introduced in error. These are termed 'insertion errors'. Increasing the word insertion penalty biases the decoder against including additional words in the output, and can help reduce the number of insertion errors.

General comments

It should be noted that to obtain new recognition scores requires that the whole process of training the language or acoustic models be repeated, the decoder re-run and a new recognition score derived. However, it is anticipated that by this point the process of doing so will have been automated through the writing of scripts, so it is expected that, beyond the actual modification of the models, the re-running of the training processes will be largely automatic processes.

Running the recognition stage on live speech input

The Sphinx 3 decoder can be set to run on live speech with no further integration or programming work needed. If any work is required in this section, it is anticipated that it will be in reformatting the output of the decoder, and piping or redirecting its output into additional programs.

It is anticipated that the main challenge of this stage of development (and indeed one of the main challenges in this development in general) will be in integrating the recogniser with a practical system. In particular, unlike the Festival speech synthesis system described in Appendix I1, Sphinx is not well-integrated with Microsoft Windows. Significant development work would have to be undertaken to ensure this integration, and it should be emphasised that this integration is not a trivial task.

Adapting the acoustic and language models given the speech input data

Standard techniques exist for acoustic model adaptation in speech recognition. The most common ones of these are MAP and MLLR. Sphinx supports both MAP and MLLR re-estimation. In particular, it allows MLLR adaptation to be performed on-line, while recognition is still taking place.

Techniques also exist for the adaptation of language models. This adaptation is beyond the scope of any initial development for Welsh, as it relies on the language model being updated off-line, and replacing the existing language model in the recognition stage once it has been updated.

General comments

It is evident from the above that the challenge of speech recognition for Welsh is a tractable and achievable one. However, it should also be implicit that previous expertise in the field is important in developing this complex area of SALT. This points towards the preservation of the existing skill base in Wales, and nurturing of future developers, as being essential in ensuring that this key technology is developed, and continues to be developed in future.

I3. UIMA: Intelligent Web Search

Retrieving information and knowledge from document repositories and/or the world wide web through search remains difficult and limited.

Simple keyword search, the simplest and most widely used method to date, asks users to enter words or terms into a text field. An engine searches through a list of documents it has indexed and returns a list of those containing the search string. This method is inefficient and inaccurate for a number of reasons. But primarily:

- A term may have various meanings and contexts in the documents identified in search results. Many of which may not be relevant to the information the user is seeking.
- The relevance of search engine results relies on the user searching with the correct term in order to retrieve relevant documents.

Attainment of new knowledge is difficult when the key pieces of information are dispersed through inefficient search results.

More intelligent searches can be achieved if search engines are able to access and utilise better analysis of the contents of documents. Such analysis is able to recognise and provide key pieces of information and data that can easily convey to a computer meanings and relationships to other data.

A number of open source projects, open web based services and open standards are in development within the Natural Language Processing and Semantic Web communities in order to facilitate improvements for search in such a way. Of particular interest are:

- The General Architecture for Text Engineering (GATE) by Sheffield University Natural Language Processing Group – see http://gate.ac.uk
- Unstructured Information Management Architecture (UIMA) by IBM and the Apache Foundation. See http://incubator.apache.org/uima/
- OpenCalais by Reuters – see http://www.opencalais.com
- Content Analysis Web Services - Term Extraction API by Yahoo – see http://developer.yahoo.com/search/content/V1/termExtraction.html

Both UIMA and GATE are open source projects and are implemented in Java. OpenCalais and Yahoo Term Extraction are services hosted and developed by the large businesses. UIMA and GATE however provide the greatest level of flexibility and freedom to Welsh SMEs for innovating new applications and services

Since publishing its source code in 2006, UIMA has been used by commercial vendors as a platform for the development and distribution of unstructured information and text analysis solutions and standardisation efforts are underway at OASIS.

In further support of UIMA, the Language Technologies Institute at Carnegie Mellon hosts a repository of UIMA natural language components and wrappers for various languages. Wrappers exist for pre-existing OpenNLP and GATE components. The

Institute has become actively involved the adaptation of UIMA for other applications such as multi-engine machine translation and large scale annotation for question answering.

The UIMA Framework in Brief

The UIMA documentation describes itself as a framework and SDK for developing applications that analyze large volumes of unstructured information such as text, audio and video, in order to discover knowledge that is relevant to an end user.

UIMA is licensed according to a BSD like license, meaning any SME or organisation may freely develop further solutions and relicense their derivative work to a license of their choice – e.g. A commercial license or even a GNU GPL license.

A very simple application may be to ingest plain text and identify entities such as persons, places, organisations and relations such as 'work-for' or 'located-at'.

Such functionality is delivered by UIMA by means of components with specific tasks and data flows that are managed by UIMA. An example series of interacting UIMA components could be :
Language Identification
→ Language Specific Segmentation
→ Sentence Boundary Detection
→Entity Detection (person/place/name etc)

The UIMA framework is able to scale to very large volumes by replicating processing pipelines over a cluster of networked nodes.

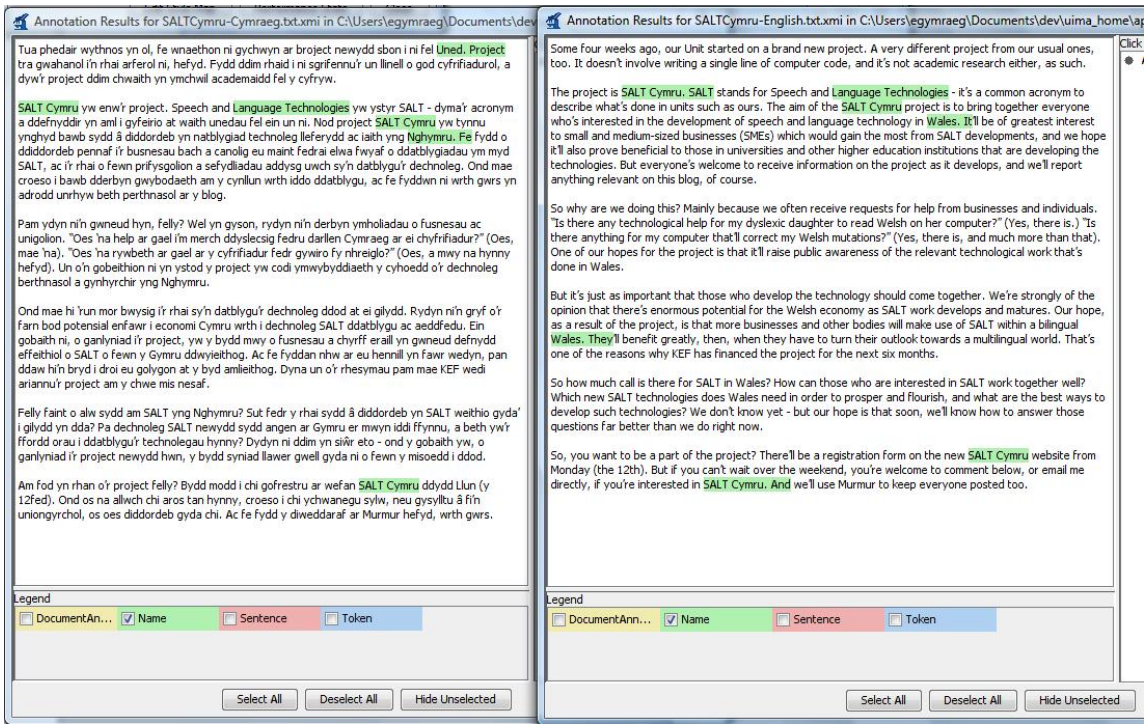
Evaluation of the UIMA Platform

The UIMA framework can be downloaded from the Apache web site at:

http://incubator.apache.org/projects/uima.html

The installation procedure is not trivial. UIMA requires the most recent versions of the Java Developers Kit and Java Runtime Environment in order to run. Installation depends on the user unzipping the download contents file and configuring various environment variables to point at the new installation. Some technical knowledge of Java and of your computer's shell or command line environment is needed to further ensure a successful complete installation of UIMA.

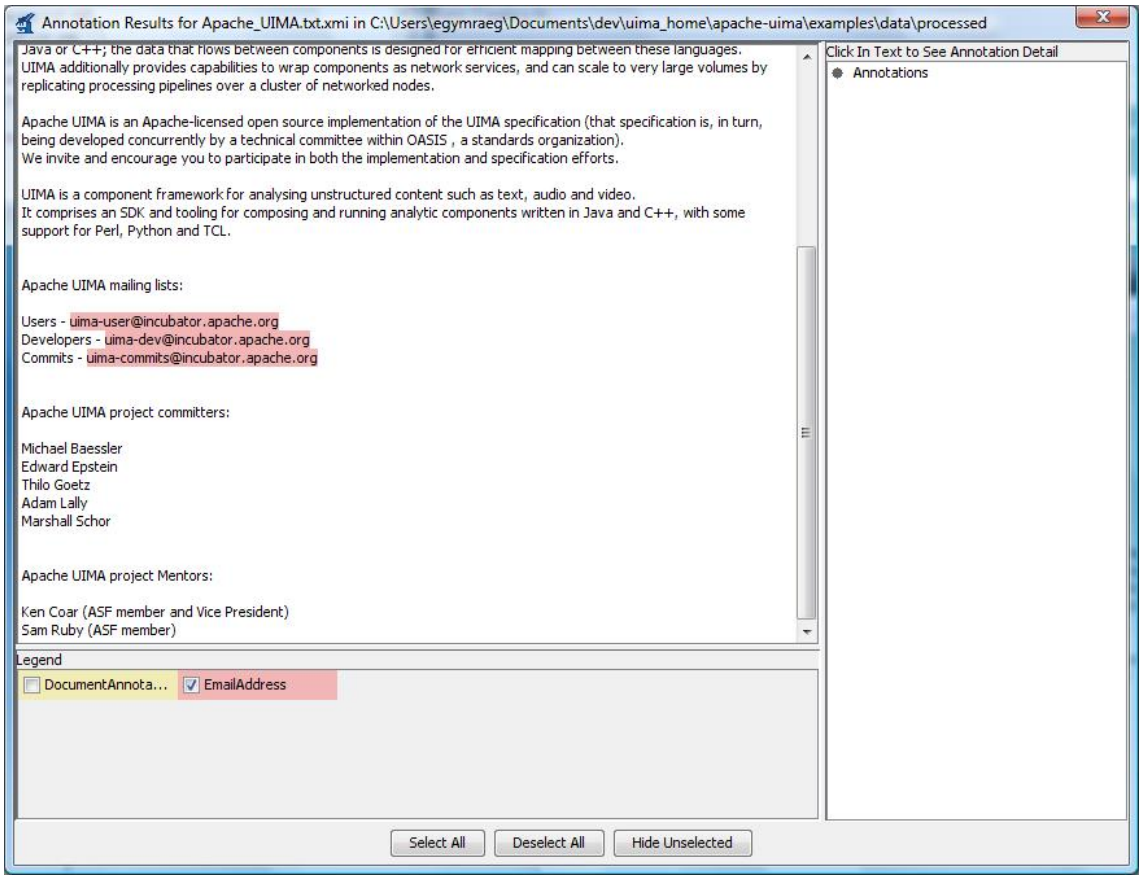
The UIMA download supports the user in getting started quickly and contains examples. Running simple example component and pipelines over any text are quite easy. The below screenshots show UIMA having annotated automatically both English and Welsh texts for names.



Detected names are highlighted in green.

Other analysis engines provided as examples by UIMA include a simple email recognizer, person names and title recognizer and a government official recognizer.

The following screenshot demonstrates UIMA's ability to recognize e-mail addresses (highlighted in red)



The examples demonstrate that some errors may occur. For example, in English texts some detected names roll over a sentence boundary (– 'SALT Cymru. SALT ...') The same level of success can be seen in the Welsh language text though this may be obvious from using components not intended for the Welsh language. Only English names are successfully recognised also in the English text are correct. Some names such as 'Cymru' are not identified at all. That is both 'Cymru' and 'Nghymru' have not been recognised.

Development Opportunities

Opportunities exist for developing and participating in the Apache UIMA. More sophisticated language processing UIMA components exist at the Carnegie Mellon UIMA Component Repository for improved accuracy in English and other source language texts.

Improving the analysis and management of Welsh language unstructured information would require the development of new UIMA compliant language processing components, dictionaries and gazetteers. Components such as lemmatizers, tokenizers, part of speech taggers, lexicons and gazetteers exist at the Language Technologies Unit that are utilised in the Unit's Welsh language proofing tools such as its spelling and grammar checking applications.

Developing UIMA framework components and UIMA based applications is made more feasible through integration in the Eclipse IDE (a software application development environment). A significant community, amounts of source code and documentation also exist for aiding the developer.

Potential Worth To Welsh SMEs

In contrast to structured information that exists in databases, indexes and knowledge bases, unstructured information is the largest and the fastest growing source of information to business and organisations. Such sources include texts from websites, blogs, news, emails, social networks, forums as well as static web documents. In recent years both audio and video have increasingly become sources of information.

Information in these sources are of high value and are expressed in human language. Thus search and other applications with the content of such sources have been inefficient and/or unrealistic.

The utilisation of the UIMA framework as a foundation to any application that deals with unstructured information present considerable opportunities for Welsh SMEs to scale up in order to manage, search and gaining knowledge from either their own internal information sources or those on the web.

The additional challenge in Wales, as in any multilingual nations, is that of seeking knowledge from information sources in multiple languages. As UIMA is increasingly used in bilingual and multilingual contexts Welsh SMEs are increasingly able to

- gain knowledge from cross-lingual sources.
- Automatically mark up their own web content with semantic meta data tags (such as microformats and W3C's rdfa) that are recognised by leading search engines such as Yahoo and Google in their continuing embracing of semantic web standards.
- With participation in the semantic web, with meanings conveyed to other computers and systems, regardless of the initial human language, information and data from Welsh SMEs will be more accessible and interoperable with the continued evolution of the web into mashups/Web 3.0 etc.

I4. Tesseract OCR Engine – a report on its potential

What is OCR technology?

OCR technology allows the conversion of scanned images of printed text or symbols (such as a page from a book) into text or information that can be understood or edited using a computer program. The most familiar example is the ability to scan a paper document into a computer where it can then be edited in popular word processors such as Microsoft Word. However, there are many other uses for OCR technology, including as a component of larger systems which require recognition capability, such as the number plate recognition systems, or as tools involved in creating resources for SALT development from print based texts.

Availability

General Availability

Commercial OCR technologies, of which OCR engines is the core component, are widely available. These commercial engines are highly developed and offer considerable accuracy when working with texts from major languages. With English text for example, the top commercial engines have an accuracy of over 98%. Some companies specializing in OCR technologies offer software developer kits (SDKs) which allow software developers to license the use of the OCR technology in their own systems.

Language Availability

As previously mentioned, the accuracy of major-language commercial OCR is very high. This accuracy is achieved through the combination of language independent algorithms for identifying the likely value of a character with language specific information such as wordlists that improve the results of these algorithms.

Commercial OCR technologies rarely include language specific information for less spoken languages such as Welsh. By attempting to identify individual characters these OCR technologies will still work to some degree with these languages, but the lack of language specific information to compare these results to leads to a considerable drop in accuracy. Unfamiliar or undefined characters such as Welsh's *ŵ* and *ŷ* may not be correctly recognized at all. In lengthy texts, the post-editing required after using OCR technologies may be tedious and time-consuming.

Price

Commercial OCR software is expensive. Market leader Nuance's Omnipage 16 currently retails for £80 per copy for the standard version, and £292 per copy for the Pro version. The company also offers licences to use their OCR engine in the form of an Omnipage SDK, the price of which can be prohibitive for SMEs involved in smaller projects.

Use

Home and Office

OCR technology is commonly used in home and office environments, where the ability to convert printed paper documents into editable electronic documents is a considerable time saver when the only alternative is to redraw or retype a document in its entirety.

Accessibility
The digitization of printed documents can be of enormous benefit to visually impaired users by enabling printed texts to be digitized and read out loud using text-to-speech technologies.

OCR Components in larger systems
OCR engines are often found as components of larger systems that are designed to track information using visual cues that have been placed on objects. An example of this is the technology used to identify the number plates of cars entering and leaving congestion zones. Similarly, OCR technology can also be used track the progress of a delivery or the progress of a component through a supply chain.

Creating Corpora and Lexica
OCR technology is also invaluable to developers that are involved in the creation of resources used by speech and language technologies. By digitizing print-based texts, developers can create electronic resources such as corpora and lexica for languages where existing digital texts are insufficient, unsuitable or do not exist. It is from corpora and lexica that resources such as word lists and grammar rules are generated. These resources lie at the heart of SALT development for any particular language.

Creating Translation Memory from printed texts
OCR technology can be of great benefit to translators as they move over to using Computer Assisted Translation (CAT), as it allows the creation of valuable translation memories from previous translations which were archived in paper form. The use of such translation memories can increase translator productivity by up to 40%.

Why is OCR of interest to SALT Cymru?

Pre-competitive advantage
The development and refinement of open source OCR technology would enable developers to flexibly and cheaply incorporate OCR technology into their systems without the burden of developing or licensing the underlying technology. By lessening the overheads involved in the development of such systems, smaller sized enterprises such as SMEs could consider moving into markets where previously only larger companies were able to compete, especially if the relevant training was also made available.

Language Support for Welsh
As mentioned previously, highly-developed OCR engines tend to only be available for major languages. This means that most of the world's languages are currently not well supported, providing an opportunity for companies wishing to specialize in providing support for these unsupported languages.

There is, for instance, currently no OCR technology in existence that produces satisfactory results when scanning Welsh and bilingual Welsh/English printed text. This is a major problem for those wishing to digitize printed texts that contain Welsh or a combination of Welsh and English (see the PowerPoint presentation given by representatives of the National Library of Wales at the JISC Digitization Conference, 2007. Link:
www.jisc.ac.uk/media/documents/programmes/digitisation/jiscdigicon07locock.ppt). The

ability to accurately digitize Welsh language texts would be of great benefit to many sectors and would enable:

- The ability to create Welsh digital language resources such as lexica and corpora for Speech and Language Technologies developers from printed resources
- Easier digitization of historical Welsh texts as undertaken by the National Library of Wales (cf. projects such as Culturenet's Books from the Past)
- The ability to process forms returned in Welsh, such as those from the Welsh Assembly Government and other public bodies
- The ability to enable blind users of Welsh text-to-speech to have access to books not available in digital form
- The creation of Welsh/English Translation Memories from existing parallel translations that survive only in printed form

Expertise developed in the process of developing OCR tools for the Welsh language could be put to commercial use with other languages that lack full OCR support. A long tail of the world's languages are in a similar position to that of Welsh.

Tesseract OCR Engine

What is Tesseract?

Tesseract is an open source optical character recognition (OCR) engine originally developed at Hewlett-Packard between 1985 and 1995, but never commercially exploited. It rated highly at The Fourth Annual Test of OCR Accuracy (http://www.isri.unlv.edu/downloads/AT-1995.pdf) held in 1995 at the University of Nevada, Las Vegas' Information Science Research Institute (ISRI: http://www.isri.unlv.edu/). However by that time, Tesseract's development had ceased.

In 2005, HP transferred Tesseract's unaltered code to the ISRI and it was released as open source. ISRI discovered that the original developer, Ray Smith (see http://research.google.com/pubs/author4479.html), was now employed at Google after several years working on the market leading commercial OCR engine *Omnipage*. Google were persuaded by ISRI to allow Smith to continue development of Tesseract as open source software. Version 2.0 is now available for download from Google Code at http://code.google.com/p/tesseract-ocr/.

Limitations of Tesseract

Tesseract is an OCR engine, not a complete OCR program

Tesseract is an OCR engine rather than a fully featured program similar to commercial OCR software such as Nuance's Omnipage. It was originally intended to serve as a component part of other programs or systems. Although Tesseract works from the command line, to be usable by the average user the engine must be integrated into other programs or interfaces, such as FreeOCR.net, WeOCR or OCRpous. Without

integration into programs such as these, Tesseract has no page layout analysis, no output formatting and no graphical user interface (GUI).

OCRopus
OCRopus is an open source document analysis and OCR system also funded by Google. It provides much of the layout analysis functionality missing from Tesseract. It is also able to use engines other than Tesseract. See: http://code.google.com/p/ocropus/.

WeOCR
WeOCR is a platform for Web-enabled OCR, which provides users with an online interface for OCR engines, including Tesseract, which allows users to upload images of English text in bmp, jpeg and pbm/pgm/ppm formats and receive the output in a text file format. It can be accessed from the following link:
http://asv.aso.ecei.tohoku.ac.jp/tesseract/ .

FreeOCR.net
FreeOCR.net is a simple but effective freeware program that uses Tesseract as its OCR engine and produces accurate results from print, via your scanner to text format when scanning English texts. It does however lack layout analysis and output formatting, and although available as freeware, FreeOCR.net is not open source. Nevertheless, it serves as an impressive and foolproof demonstration of the potential of the Tesseract engine. It can be downloaded from: http://softi.co.uk/freeocr.htm.

Unsupported features
Although Tesseract has been modified to deal with UTF-8 characters, Tesseract may not work well with languages that possess complex characters, or connected scripts such as Arabic. Only left-to-right scripts are supported. Right-to-left texts are currently processed as if they were as if they were left-to-right texts. ASCII punctuation and digits are expected by the code, so any language using alternatives to these will not be fully supported.

How does Tesseract work?

A comprehensive overview of the Tesseract OCR Engine entitled *An Overview of the Tesseract OCR Engine* by Ray Smith is available from the IEEE, at the following address:

http://ieeexplore.ieee.org/iel5/4376968/4376969/04376991.pdf?tp=&isnumber=4376969&arnumber=4376991 (Subscription or payment may be required)

For convenience, the following is a brief overview of how Tesseract works:

1. Outlines are analysed and stored
2. Outlines are gathered together as *Blobs*
3. Blobs are organized into text lines
4. Text lines are broken into words
5. First pass of recognition process attempts to recognize each word in turn
6. Satisfactory words passed to adaptive trainer

7. Lessons learned by adaptive trainer employed in a second pass, which attempts recognize the words that were not recognized satisfactorily in the first pass
8. Fuzzy spaces resolved and text checked for small caps
9. Digital texts are outputted

During these processes, Tesseract uses:

- algorithms for detecting text lines from a skewed page
- algorithms for detecting proportional and non proportional words (a proportional word is a word where all the letters are the same width)
- algorithms for chopping joined characters and for associating broken characters
- linguistic analysis to identify the most likely word formed by a cluster of characters
- two character classifiers: a static classifier, and an adaptive classifier which employs training data, and which is better at distinguishing between upper and lower case letters

The project page for Tesseract is located on Google code at the following address: http://code.google.com/p/tesseract-ocr/

This page features background, details of important changes and supported platforms in addition to a project roadmap and a list of developers. Downloads of the Tesseract engine, as well as associated files and utilities are also located her, and an associated Google Group can be found at http://groups.google.com/group/tesseract-ocr.

How accurate is Tesseract OCR?

The above processes ensure that Tesseract is highly accurate when recognizing texts from languages that are currently supported. Results from The Fourth Annual Test of OCR Accuracy (http://www.isri.unlv.edu/downloads/AT-1995.pdf) are still available online, where, for example, Tesseract demonstrated a Word Accuracy of 97.69% with a sample of English newspapers. Since these tests, the Tesseract development team at Google claim to have improved Tesseract's general results by 7.31% for Tesseract version 2.0.

Above: a real world example of the Tesseract OCR engine in action, using FreeOCR.net

What are the language-specific components of Tesseract?

For a language such as English, 8 components are used:

1. General Words Wordlist (tessdata/eng.word-dawg)
2. Frequent Word Wordlist (tessdata/eng.freq-dawg)
3. User Wordlist (tessdata/eng.user-words)
4. Index for Character Set (tessdata/eng.inttemp)

5. Box file – for use in locating characters in the training file (tessdata/eng.normproto)
6. Box file – for use in locating characters in the training file (tessdata/eng.pffmtable)
7. Language's Character Set (tessdata/eng.unicharset)
8. Character Cluster Disambiguator - for 'm' and 'rn', for instance. (tessdata/eng.DangAmbigs)

OCR technology uses character recognition to attempt to identify the individual characters that make up a printed text. Although the process used to identify individual characters is language independent, Tesseract must be given a list of the specific characters used by a language (item 4 in the list above).

Tesseract must then be trained to correctly identify these characters when they appear within a piece of text. Training is done by feeding into Tesseract a document with words, sentences, symbols and numbers from the required language which contains a recommend ten to twenty example of each of the characters used by that language. Such a list has been added to this document as an appendix. This list must be fed in twice, once as digital text and once as a scan of a printed version of the same text. This produces a 'boxfile' containing Tesseract's interpretation of the position of characters and their identity.

The next part of the process is to manually correct any errors made by Tesseract, for example the identification of ŵ as W or the identification of the letter combination *rn* as *m*. A useful utility with a graphical user interface now exists to simplify this process, and is available from the Tesseract project page. Once this task has been finished, common mistakes such as those mentioned above can be added to the Character Cluster Disambiguator file. This training process must be repeated with all font types required, including bold, italic and underlined versions of the same font. The Character Cluster Disambiguator file, in conjunction with a language's word list, helps Tesseract identify a word by suggesting possible corrections to certain characters that allow Tesseract to locate the correct word in its word list. For example, the file can be used to suggest to Tesseract that *rn*, *wr*, *iii*, and *an* could all potentially be misidentifications of the letter *m*, and Tesseract will search the wordlist accordingly.

However, not all languages will have a list of the commonly used words at their disposal. A list of the head words from a dictionary, for example, is not sufficient as all inflected forms must also be included. For example, *mouse* and *mice* should both be included in an English wordlist, and so too *run* and *ran*. Many other languages undergo far more inflection than English, so their corresponding wordlists are likely to be both longer and harder to create. In Welsh for example, nouns like *coffi* (coffee) occur regularly as *goffi*, *choffi* and *choffi,* effectively quadrupling the number of nouns in a list. Many European languages have significantly more verbal forms compared with English. This inherent complexity in language is part of the reason that resources such as wordlists have not been develop for many languages with less resources. Bespoke wordlists would have to be created for any language supported where wordlists are not available. In truth, for optimum performance, Tesseract requires not one, but two word lists. One should contain the most frequently used words in a language, which Tesseract will search first, the second, which Tesseract will only search after failing to find a word in the first list, should contain the less frequently used words in a language. A third list for user-added words also exists.

In theory, the above steps should allow for the creation of an OCR engine in languages currently unsupported by Tesseract. However, some languages may not be suitable candidates, as right to left languages are currently not compatible with some of

the hardcoded functionality built into Tesseract. Depending on character sets, some languages with complicated glyphs or characters may also be unsuitable. However, Google are currently working on increased language support in future versions of Tesseract.

Development

Tesseract development is currently being led by Google, under the direction of Ray Smith, Tesseract's original developer and one of the foremost experts on OCR technology. Although the participation of Google is of obvious benefit to the project, detailed information concerning the exact nature of the work being undertaken by Google's software engineers between updates is not made publicly available. It is therefore difficult for independent developers to coordinate their work with that being done by Google, and presently developers cannot be certain that their work will be compatible with, and not duplicate, work already carried out by Google.

An example relevant to language development is the following statement made by the lead developer, Ray Smith, at Google:

> "Although it is very tempting to try to expand tesseract to new languages, if you did so, you would be overlapping significantly with the work going on at Google. Of course that leaves anyone that wants a different language in the difficult position of either waiting for it to be available, or trying to train it themselves. I will be in a much better position to discuss language compatibility after the next release, by which time there will be much more language support."

It would therefore seem a sensible precaution for anyone intending to develop the language aspect of Tesseract to first attempt to liaise with the developers at Google.

Tesseract Roadmap
(This roadmap is taken directly from Tesseract's page on Google Code. Accessed 29/03/08.)

Version 2.00 is now available and contains the following new features:

- Support for English, French, Italian, German, Spanish, Dutch
- Scripts to test accuracy against the original 1995 tests run by UNLV
- Ability to train in other languages and scripts

We are considering the following features for upcoming releases:

- ground truth data release
- integration with OCRopus (http://www.ocropus.org/), to support layout analysis
- integration with Leptonica (http://www.leptonica.com/), to support layout analysis and more image formats
- support for even more languages
- high-resolution character shape modelling for improved recognition rates
- a GUI frontend (again, probably shared with OCRopus)

Licence

Originally developed by Hewlett-Packard, Tesseract was released under the Apache 2.0 open source licence ten years after its development had come to an end. The Apache 2.0 open source licence is considered a free software license, compatible with version 3 of the GPL, by the Free Software Foundation. It allows the freedom to use the software for any purpose, including its distribution, its modification, and the distribution of modified versions of the software. However, unlike LGPL licences, the Apache 2.0 license does not require modified versions of the software to be distributed under the same license as the original software. This allows the direct commercial exploitation of modified versions, making the software attractive to business whilst at the same time not undermining the usefulness of the original software to those wishing to develop open source products.

Appendix J. An examination of open innovation

High Impact University Business, Enterprise and Innovation: Exploring Opportunities to Create Knowledge-based Businesses

Reading University, 18th December 2007

This was a one day conference sponsored and organised by the key stakeholders in the newly created High Impact University Business and Industry (HIUBI) community. The HIUBI community aims to bring about synergies, opportunities and solutions to the UK University business, enterprise and innovation space.

The main catalysts for the new community are:

- Microsoft (who hosted the event at the UK offices in Reading)
- National Council for Graduate Entrepreneurship (NCGE)
- UNICO – representing Technology Exploitation companies of UK Universities
- Association for University Research and Industry Links (AURIL)
- Institute of Knowledge Transfer (IKT)
- UK Business Incubation (UKBI)
- The Higher Education Academy

As well as launching the HUIBI community website at http://hiubi.ncge.com the day was dedicated to discussing how all participants can continue to work together through four action tracks:

- Action Track 1 – facilitating interaction between university professionals within university internal university communities.
- Action Track 2 – opportunities to increase, facilitate and enhance University practitioner and professional self learning
- Action Track 3 – measure the impact of university interaction and interface with business and society.
- Action Track 4 – improving the external interface between Universities and business and society that will result in the creation of more knowledge based businesses.

Morning Session

The day began with motivation for increasing opportunities for University and Business collaboration and knowledge transfer and improving their effectiveness.

It was stated that knowledge based companies constitute 40% of the UK economy. This share is destined to grow to 50% by 2010. Universities and knowledge transfer will be a key in delivering this growth.

It was also suggested that further opportunities for collaboration may lie from realising that approximately 80% of UK businesses in the service sector have yet to work with a University. Here service and technology innovations will be crucial.

Further opportunities exist from multi-sector collaboration where new and non-obvious scenarios could benefit from knowledge transfer, e.g. the use of SALT by an electrical equipment home installation company, or in agricultural scenarios or by the creative industries.

As opportunities in collaboration increase in U2B, so will its nature and culture of innovation change into be more distributed in order to be able to scale up and deliver more opportunities. The change will take time but may eventually mimic or follow models labelled as 'Open Innovation' or 'Distributed Innovation' which states that all organisations benefit and advance their own innovations from having purposeful inflow and outflow of knowledge. This is in contrast to the traditional innovation model which was more closed in nature, and organisations preferred to keep their discoveries confidential from all external organisations, and relied on their own research and development. The more open innovation model recognises that the smartest people and the most exciting innovation may well work and derive value for other organisations but that some internal R&D can be conducted to claim some portion of that value.

It was widely recognised between delegates that increasingly that the smartest innovation may well be distributed outside the UK and in other countries. As UK universities should look at benefiting from opportunities at collaborating with international scholars and companies in whichever part of the globe, be it in the US, the EU or upcoming developing countries such as India or China.

The open innovation models are encouraged by the EU as well as companies such as IBM, Proctor and Gamble and Microsoft.

In such a culture it is recognised that knowledge is shared and transfer becomes a two way flow. The collaboration can take many forms such as research partnerships, collocation, consultancy, licensing etc. Relationships and collaboration operate on higher levels of trust and are based upon initial and continued face to face network building activities in clubs, special interest groups, associations or online networks.

Trust needs to be established from the beginning and both parties are encouraged to talk about fears and worries from collaborations as well as their solutions. Preconceptions are to be given a chance to be eliminated. In all collaborations, each parties goals are respectable, even those such as earning money. One party should not aim to dominate the collaboration with their aims.

It was further noted that Universities have their students and graduates (through alumni offices) as a captive audience for promoting participation their University and Business collaborations opportunities. If the students prove themselves to be of calibre in collaborations then they become ambassadors for the University.

In measuring the success of Universities and Businesses collaboration, it was stated that the current metrics are not sufficient and ineffective for current and future support of collaborations. Current metrics are too revenue based. Value exists in policy changes, people flows etc not just in economic contexts and therefore it was suggested that new more intelligent metrics are developed to measure

- reputation
- visibility, accessibility and delivery of research.
- Finding new and retaining existing collaborating businesses.

The impacts of collaborations are hard to quantify since the benefits are often long term. Universities wish to have their collaboration in the context of social enterprise such as regeneration, community engagement, linguistic and cultural diversity support which are hard to quantify.

The morning session closed by mentioning that such collaboration networks will increasingly require investment in people as well as in research in new technology. In this, the challenge is for modernising the academic workforce so that

- academics not only research (always their preferred area of work) but also participate in development projects.
- ultimately every academic becomes a knowledge transfer professional.