MACHINE TRANSLATION AND WELSH: THE WAY FORWARD

CYFIEITHU PEIRIANYDDOL A'R GYMRAEG: Y FFORDD YMLAEN

Harold Somers

Centre for Computational Linguistics

UMIST

Manchester

A Report for The Welsh Language Board

July 2004

CONTENTS

EXECUTIVE SUMMARY

The state-of-the-art of Machine Translation (MT) and the needs of the Welsh-speaking community indicate that it is timely to invest in the development of language technology provision for Welsh, in particular MT.

Despite 50 or 60 years of research, and some major advances, MT technology can not yet offer translation quality to match that of human translators. Nevertheless, usage of free on-line MT systems by WWW users indicates that for certain types of text, MT can provide a useful and usable level of rough translation. Furthermore, in combination with controls on the language used for example in technical authoring, MT can provide an excellent quality of first-draft translation needing little revision, and offering great savings in translation costs.

This report gives as background the history of MT research, and explains why translation is so difficult for a computer. It is noted that there are currently three different technological approaches to MT – the traditional "rule-based" approach, a wholly statistical "machine learning" approach, and a kind of hybrid, generally called "example-based" MT – and identifies the underlying needs in terms of resources for these approaches. We then discuss in detail the level of existing provision for Welsh language technology, and the availability of resources for each of these approaches.

It was felt to be useful to compare the Welsh situation with some other "minority" languages, and the cases of Irish, Catalan, Basque and Galician are discussed in detail. Maltese and the other Celtic languages are also briefly discussed. These existing cases inform us in a variety of ways, but we are lead to conclude that actually the situation for Welsh is one of the most favourable of the minority languages.

The report then outlines a proposal to fund a two-year project to develop in parallel three types of MT system: a traditional rule-based Welsh–English system, a statistics-based system initially Welsh–English, then English–Welsh, and an example-based system, initially English–Welsh, to be adapted to Welsh–English. A project management role is also identified, including regular contact with the sponsors, and two project workshops (one at start-up, one at the end) are also recommended.

LIST OF ABBREVIATIONS

ALPAC	Automated Language Processing Advisory Committee
BIML	British indigenous minority language
CAT	Computer-assisted translation
CEG	Cronfa Electroneg o Gymraeg
CRPIH	Centro Ramón Piñeiro para a Investigación en Humanidades
DCU	Dublin City University
EBMT	example-based machine translation
EPSRC	Engineering and Physical Sciences Research Council
ESRC	Economic and Social Research Council
FAHQT	fully automatic high quality translation
IT	Information technology
ITE	Institiúid Teangeolaíochta Éireann
LIP	Language interface pack
LT	Language Technology
MT	Machine Translation
NLP	natural language processing
POS	part of speech
RBMT	rule-based machine translation
SL	source language
SLT	spoken-language translation
SMT	statistical machine translation
TL	target language
TM	translation memory

1 INTRODUCTION

This document is a report on research to develop part of a strategy for Language Technology (LT) and the Welsh language, focussing on the provision of Machine Translation (MT) between Welsh and English, in both directions.

This report consists of three main parts:

In Section 2 we review the history of MT, how MT works, its limitations and prospects, all with special reference to the case of "minority languages" in general, and Welsh in particular.

In Section 3 we review the current situation regarding LT provision for Welsh, and the perceived needs for English–Welsh MT. And we review LT and MT in some other minority-language countries

In the final section we report on the practical prospects for the development of MT systems for Welsh, including recommendations for funding levels and an identification, where appropriate, of individuals and organizations who are in a good position to implement the recommendations.

A note on terminology: referring to Welsh as a "minority" language is consistent with the fact that it is spoken by a minority, albeit significant, of the population. No other inference should be drawn from the use of this term, in particular regarding the appropriateness of Welsh for any linguistic task.

The term "machine translation" is the most widely used English term to refer to the process (or result) of using a computer to produce, automatically or largely automatically, a translation. The term, felt by many to be inappropriate or misleading, has a historical basis (who calls a computer a "machine" these days?), and indeed was at one time in competition with the alternative "automatic translation", the latter still preferred in French (*traduction automatique*) and Russian (*автоматический перевод*) among other languages.[1] A wider term, "computer-aided" or "computer-assisted" translation (CAT) refers to computer-based software tools aiming, as its name suggests, to help a human – usually a translator – to produce a translation. Indeed, partial MT may be just one of several tools in a CAT package.

[1] It is interesting to note that in coining the Welsh term *Cyfieithu Peirianyddol*, the English model has been followed, while *Cyfieithu Cyfrifiadurol* or *Cyfieithu Awtomatig* might also have been possible.

2 WHAT IS MACHINE TRANSLATION?

2.1 BRIEF HISTORY OF MT

A mechanical translation tool has been the stuff of dreams for many years.[2] Translation has been a suggested use of computers ever since they were invented (and even before, curiously). The history of MT is usually said to date from a period just after the Second World War during which computers had been used for code-breaking, based on which Warren Weaver, then vice-president of the Rockefeller Foundation, tried to raise interest among numerous colleagues in using the new digital computers for translation. There was a mixed reaction to Weaver's ideas, but between 1950 and 1965 MT research groups worked in a number of countries, particularly the USA, funded by government, military and private money – at least $12m and perhaps as much as $20m.

In 1964, the US government set up the Automated Language Processing Advisory Committee (ALPAC) to see if its money had been well spent. Their report, published in 1966, was highly negative about MT with very damaging consequences. Focussing on Russian–English MT in the USA, it concluded that MT was slower, less accurate and twice as expensive as human translation, for which there was in any case *not* a huge demand. It concluded, infamously, that there was "no immediate or predictable prospect of useful machine translation". In fact, the ALPAC report went on to propose instead fundamental research in computational linguistics and NLP (natural language processing), and suggested that machine-*aided* translation may be feasible. The damage was done however, and MT research declined quickly, not only in the USA but elsewhere.

Actually, the conclusions of the ALPAC report should not have been a great surprise. Early efforts were hampered by primitive technology, and a basic under-estimation of the difficulty of the problem on the part of the researchers, who were mostly mathematicians and electrical engineers, rather than linguists. It is at about this time too that much repeated (though almost certainly apocryphal) stories about bad computer-generated translations became widespread. Reports of systems translating 'out of sight, out of mind' into the Russian equivalent of 'blind idiot', or 'The spirit is willing but the flesh is weak' into 'The vodka is good but the meat is rotten' can be found in articles about MT in the late 1950s though looking at the systems that were around at this period one has difficulty in imagining any of them able to make this kind of quite sophisticated mistranslation, and it has been suggested that similar stories have been told about incompetent *human* translators.

Despite the ALPAC report, the 1970s and early 1980s saw continuing MT research in Canada, Western Europe and Japan, and even, in a much more limited way, in the USA (privately funded). Research was fuelled by multilingual policies in Canada and in the European Communities while in Japan, some success with getting computers to handle the complex writing system of Japanese had encouraged university and industrial research groups to investigate Japanese–English translation.

Systems developed during this period largely share a common design basis, incorporating ideas from structural linguistics and computer science. System design divided the translation problem into manageable subproblems – analysing the input text into a linguistic representation, adapting the source-language (SL) representation to the target language (TL), then generating the TL text. The software for each of these steps would be separated and modularised, and would consist of grammars developed by linguists using formalisms from theoretical linguists rather than low-level computer programs. The

[2] This section is based on the Introduction to the present author's book *Computers and Translation: A Translator's Guide*, Amsterdam (2003): Benjamins.

dictionaries likewise were coded separately in a transparent manner, so that ordinary linguists and translators could work on the projects, not needing to know too much about how the computer programs actually worked. This "second-generation" approach to MT is now generally known as the "rule-based" approach, in contrast with alternatives that appeared later.

By the mid 1980s, it was generally recognised that fully automatic high-quality translation of unrestricted texts (FAHQT) was not a readily achievable goal for the near future, and researchers started to look at ways in which usable and useful MT systems could be developed even if they fell short of this. We now distinguish between the use of MT for *assimilation*, where the user is a reader of a text written in an unfamiliar language, and *dissemination*, where the user is the author of a text to be published in one or more languages. In particular, the idea that MT could work if the input text was somehow *restricted* gained currency. This view developed as the "sublanguage" approach, where MT systems would be developed with some specific application in mind, and is exemplified by the highly successful *Météo* system, developed at Montreal, which was able to translate weather bulletins from English into French, a task which human translators obviously found very tedious. Closely related to the sublanguage approach is the idea of using controlled language, as seen in technical authoring.

The other major development, also in response to the difficulty of FAHQT, further supported by the emergence of the PC, was the concept of computer-based *tools* for translators or CAT. Here, the translator would be provided with software and other computer-based facilities to assist in the task of translation, which remained under the control of the human. These tools would range in sophistication, from the (nowadays almost ubiquitous) multilingual word-processing, with spell checkers, synonym lists ("thesauri") and so on, via on-line dictionaries (mono- and multilingual) and other reference sources, to machine-aided translation systems which might perform a partial draft translation for the translator to tidy up or post-edit. As computers have become more sophisticated, other tools have been developed, most notably the Translation Memory tool now familiar to many translators.

Coming into the 1990s and the present day, we see MT and CAT products being marketed and used (and, regrettably sometimes misused) both by language professionals and by amateurs, the latter for translating e-mails and World Wide Web pages. Contemporary research since the 1990s has focused on three areas. One is the development of spoken-language translation systems, a task depending on speech recognition and synthesis as well as translation, and still in its infancy (see below). The second and third areas are inter-related and are also of much more relevance to us. This is the development of alternatives to the rule-based approach to translation, in which computers "learn" how to translate on the basis (only) of large amounts of previously translated material, with no reliance on linguistic analysis and hence on linguistic theory. This approach is seen to be particularly attractive for the rapid development of new language pairs – the third area of current focus – which typically have not had much attention from computational linguists and therefore lack the appropriate resources (computational grammars and dictionaries).

2.2 HOW MT WORKS

There are currently three main approaches to MT: the traditional "rule-based" approach, a wholly statistical "machine learning" approach, and a kind of hybrid, generally called "example-based" MT. While the three approaches are fundamentally different, a number of projects have shown that it is possible to integrate all three in a "multi-engine" system. These work either by setting all three engines to work and then reconciling the results, or else choosing the engine which *a priori* appears to be likely to give the best result. In this section, we will give a brief non-technical description of how each of these approaches works, illustrating as much as possible with hypothetical examples from Welsh–English translation.

2.2.1 WHY IS MT HARD?

Let us start by considering the so-called "first-generation" approach, perhaps more intuitive, but unsuitable for all but the simplest of translation tasks. This mainly dictionary-based approach suggests that translation is mainly about looking up words in a dictionary, then adjusting the word-order and making any other local adjustments, e.g. for number or gender agreement. For a morphologically complex language like Welsh, the inadequacy of this approach is all too easy to demonstrate. Consider the English input sentence (1).

(1) We apologize for the late arrival of the London train.

If we translate word-for-word,[3] we might produce something like (2) (as is customary in linguistics, we indicate an ungrammatical sentence with an asterisk).

(2) * Ni ymddiheuro am 'r diweddar cyrhaeddiad am 'r London trên.

Even some easily applied general rules like inverting adjective–noun sequences will only go a little way to improving this translation.

The problem is that we need to identify the correct relationships between the words, which often involves choosing amongst alternative interpretations of individual words and/or amongst competing interpretations of the whole sentence. In our example (1) we have to identify 'train' as a noun rather than a verb, and that 'late' means 'tardy' rather than 'dead'. Ambiguity problems such as these are highly significant for MT. Huge numbers of words in English are noun–verb ambiguous, sometimes with related meanings, sometimes not. Noun–adjective ambiguities are also common. Usually the surrounding context serves to disambiguate, but not always. Often, it is only "common sense" that tells you what the correct reading is.

Appropriate translation depends on the correct interpretation of these ambiguities. Humans are very good at disambiguating, and often do not even notice alternative possible interpretations. For a computer it is not so easy. Some ambiguities can be resolved by looking at the sequence of word categories. For example, the word 'book' can be a noun or a verb, but in 'the book' it must be a noun. Other times it is less clear-cut.

Even if the words and structures are correctly disambiguated, that is not the whole story. When it comes to translation, different languages express the same ideas differently. The TL may make some lexical distinctions that the SL does not, e.g. *edrych*, *craffu* and *syllu* might all be given as translations of 'look'. Often, the sentence structure of the SL is not wholly appropriate for the TL. Our example (1) illustrates this nicely: In (3) we have a grammatically correct but rather too literal translation (we prefer to call it "structure-preserving"), an improvement on (2), but still less natural than (4), which rephrases the English nominalization.[4]

(3) Ymddiheurwn am gyrraeddiad hwyr y trên o Lundain.

(4) Mae'n ddrwg gennym bod y trên o Lundain yn hwyr.

In the following sections we will see to what extent we can expect MT to do this.

[3] This translation was obtained by submitting the English words one-by-one to the BBC-Wales on-line English–Welsh dictionary http://www.bbc.co.uk/cgi-bin/wales/learnwelsh/welsh_dictionary.pl and taking the first option offered. The word 'London' was not in the dictionary.

[4] I am grateful to Gwynfor Hughes and Bryn Jones for help with the Welsh examples.

2.2.2 RULE-BASED MT (RBMT)

In RBMT, as its name suggests, we have linguistic "rules" which describe and determine how to interpret the internal structure of the words ("morphology", e.g. the '-s' of 'cakes' indicates plural), as well as how the words combine to form sentences ("syntax", e.g. article+noun is OK, but article+verb is not: in this way the category of 'book' in 'the book' is identified). Of course the rules are generally more complex than that.

Other rules may relate to more subtle aspects. In example (1), we saw how the translation of a word like 'for' can depend on its function in the sentence. Our RBMT system would encode the fact that a verb like 'apologize' expects the preposition 'for' to indicate the reason for the apology (and 'to' encodes the recipients of the apology), in contrast with phrases like 'buy a present for' where it indicates the recipient, or 'look for' where it indicates the object (cf. 'seek').

This is the traditional "second-generation" approach to MT in which input text is first analysed into a more or less abstract representation. This representation of the SL text is then manipulated so as to be more appropriate for the TL, from which the TL text is then generated. This manipulation, usually called "transfer", is also rule-based, as is the "generation" process. These "representations" are often in the form of a tree structure, very familiar to linguists, which show explicitly the relationships between the parts of the sentence, though other representations are also possible. An example of a transfer rule would be one which took a nominalization like 'late arrival of the train' and changed it to 'the train arrived late' (this rule can be generalized to cover all nominalizations). Other transfer rules might be more straightforward, the simplest of all being the ones which simply "map" SL words onto the appropriate TL words. Rules for TL generation deal only with TL phenomena such as word-order, agreement, mutation and so on.

But where do these rules come from? This is perhaps the biggest draw-back to the rule-based approach to MT, because someone has to write them! Typically this is the job of computational linguists who study the language(s) concerned and try to write computational grammars that correctly analyze and generate grammatical structures. This task is facilitated by the existence of various rule-writing "formalisms" which have been developed over the years, with associated software so that the grammars can be tested on the computer. In some cases there are even sophisticated software tools which make the job easier, by allowing the grammar writers to check their grammars for consistency and redundancy (you do not want multiple rules which cover the same construction), and to visualize with graphic interfaces how the rules work.

By far the biggest "overhead" in RBMT is the dictionary. Although we have suggested here that there is a lot more to translation than just looking up words, nevertheless dictionary look-up is of course an essential part of the process, and this means that the MT system must have access to a "machine-readable" dictionary. The information stored in an MT system's dictionary is somewhat different from what is expected of a dictionary for human use. For MT we need to know the grammatical properties of a word, some of which may be obvious to human users (e.g. we 'wait <u>for</u>' something, 'look <u>at</u>' something, 'decide <u>on</u>' something), and some aspects of its meaning have to be made explicit (e.g. that the object of 'eat' is usually something edible). Other information, such as its etymology, is irrelevant. For translation, we need to know the TL equivalent or, more usually, equivalent*s* … and how to choose between them. In bilingual dictionaries aimed at humans, this is often very vaguely expressed. For example, for 'arrival' (seen in our example sentence), the BBC-Wales on-line dictionary[5] gives simply

[5] See footnote 3.

arrival, n, cyrhaeddiad; (of person) dyfodiad

leaving it to the user to infer what the distinction is.

RBMT is a well-understood paradigm, and we can safely predict the likelihood of success depending on various factors. The most successful commercial MT systems to date are all rule-based, though many reflect a long period of development and investment. To develop an English–Welsh RBMT system would require pre-existing computational linguistic expertise and suitable lexical data (dictionaries). In other words, **assuming the ground work to have been done, we could envisage an RBMT system being developed in a reasonable amount of time**. We consider this in more detail in Section 4.

2.2.3 STATISTICS-BASED MT (SMT)

An entirely different approach to MT has been in development since the early 1990s, partly in response to the "overhead" of needing computational linguists and lexical resources to build MT systems. The basic idea of SMT is that the computer can "learn" to translate on the basis of huge amounts of data representing previous translations. Experiments were first carried out at IBM's research labs in New York state, using the parallel English–French Canadian Hansards (parliamentary proceedings).[6]

The idea is that the computer can calculate the most *probable* translation of a given input by calculating the most probable set of TL words corresponding to the words of the input sentence (the "translation model", and the most probable sequence of TL words, once chosen (the "language model"). All the probability scores are calculated in advance from the evidence of the parallel text corpus which we will call the "training data". These various probabilities interact with each other, and so a quite sophisticated mathematical model is needed to arrive at the highest-scoring combination of probabilities for a given input sentence. Let us look at these factors in a little more detail.

The translation model, i.e. the set of probabilities that a given SL word w_s is translated by a certain TL word w_t, is calculated by comparing the relative distribution of all the SL and TL words in the training data. This corpus must first be "aligned" so that each SL sentence is explicitly associated with its corresponding TL sentence(s). Then we can easily calculate the extent to which a given w_s corresponds to a given w_t: we count (*a*) the number of sentences where w_s occurs on the SL side and w_t occurs on the TL side, (*b*) the number of sentences where w_s does not occur and neither does w_t , (*c*) the number of sentences where w_s occurs in the SL sentence, but w_t does not occur in the corresponding TL sentence, and (*d*) vice versa. If $a+b$ is very high, and $c+d$ very low, this is strong evidence of a probable correspondence. If the training data contains exactly 9,000 different words in both languages,[7] this would imply 81m scores, the probabilities for each and every possible SL–TL correspondence. Most of these scores will hopefully be close to zero.

This simple picture is complicated by a number of problems. First, in general there is not a 1:1 correspondence between SL and TL words. This may be because of different inflection patterns in the two languages (for example, 'all' in English corresponds to *tout*, *tous, toute* and *toutes* in French depending on number and gender; word forms in Welsh will change with mutation). It may also be because of polysemy (our earlier example of 'look' vs. *edrych*, *craffu*, *syllu*). Another possibility is where a single word in one language is translated as a phrase in another (e.g. 'cheap' in French is *peu cher*). A second problem is that grammatical function words like 'a', 'the', 'is', occur in almost every sentence, so the co-

[6] P. Brown et al. 'A Statistical Approach to Machine Translation', *Computational Linguistics* **16** (1990), 79–85.

[7] This was the number of different word forms in the first statistical MT experiment – see footnote 6.

occurrence statistics can be misleading: if every sentence contains both 'the' and 'of', and correspondingly in French *le* and *de*, it might look like the French for 'the' is *de*. A third problem is that the sentence you are translating might contain a word that does not happen to occur in your "training" data, so the probability scores for this word are universally zero (the "sparse data" problem), which is going to mess up your calculations later. All of these problems are overcome by various "patches" to the translation model, that we need not go into here.

The language model is a little simpler. In this case we need to look only at the TL side of the training corpus. Ideally, we would like to know how probable any given sequence of words is. However, we can only go by the sequences of words that occur in the training data. Here the notion of "*n*-grams" comes in: an *n*-gram is a sequence of *n* words. We can easily count how frequent any given sequence is in our training data. For low values of *n*, such as 2 and 3, we will find that there are huge numbers of *n*-grams, with a wide range of frequencies. For example the bigram 'of the' will be very frequent, as will the trigram 'is not a'. Other sequences will be less frequent, and some such as 'the a' will not occur at all. As the value of *n* gets bigger, the frequencies will get smaller, and quickly lose their capacity to discriminate between more or less likely sequences. Furthermore, supposing again that we have 9,000 different words in our data, there are 81m different possible bigrams, 729×10^9 different possible trigrams, 6561×10^{12} 4-grams and so on. Even with the fastest computers, the number of calculations needed soon becomes impractical. Most language models find that trigram statistics are sufficient. Again we have to overcome the problem of "sparse data", this time word sequences for which we have no evidence.

The calculation of all the statistical parameters is a massive job, even for the fastest computers. But it is a job that need only be done once. Thereafter, we simply input the SL sentence, and the statistical probability calculation, although sophisticated and complex, is quickly done, and the proposal(s) for the translation produced. The attraction of the whole thing is that no knowledge of the language pair is necessary: the computer program "learns" the models automatically, and so the method can be applied to any language pair, just as long as there is a sufficient amount of aligned parallel text to train the system.

That's the theory, but does it work? Perhaps surprisingly, since the statistical process employs no linguistic insight at all, the technique can work quite well. The original experiments with English–French produced about 60% usable translations, and a lot of the errors were of a simple nature (such as wrong genders). Since then, researchers have modified the basic idea, in particular noting that a small amount of linguistic knowledge, easily obtained, can help enormously. For example, knowledge about inflection paradigms (mutation in Welsh for example) can strengthen both the language and translation models. Experiments have been carried out with an array of languages, and it seems that the technique suits some language-pairs better than others. Not surprisingly, typologically similar languages are easier to translate in this way. Some linguistic features seem to be very inhibitive (free-word-order languages are difficult, for example).

Since software packages to train an SMT system are readily available, and the only other pre-requisite is large amounts of parallel text, **it would seem to be worthwhile, and would not represent a massive investment, to experiment by developing an English–Welsh and/or Welsh–English SMT system**. A Welsh–English SMT system is reported by Phillips (2001),[8] and illustrated in Section 3.2.4 below. We consider this proposal in more detail in Section 4.

[8] J. D. Phillips, 'The Bible as a basis for machine translation', *Proceedings of PACLING, Pacific Association for Computational Linguistics*, Kitakyushu, 2001. Available at http://afnlp.org/pacling2001/pdf/phillips.pdf

2.2.4 EXAMPLE-BASED MT (EBMT)

Example-based MT (EBMT) can be seen as a hybrid of RBMT and SMT. Like SMT, it depends on a corpus of already existing translations, which it reuses as the basis for a new translation. In this respect it is similar to (and sometimes confused with) the translator's aid known as a Translation Memory (TM). Both EBMT and TM involve matching the input against a database of real examples, and identifying the closest matches. They differ in that in TM it is then up to the translator to decide what to do with the proposed matches, whereas in EBMT the automatic process continues by identifying corresponding translation fragments, and then recombining these to give the target text. The process is thus broken down into three stages: "matching" (which EBMT and TM have in common), "alignment", and "recombination". Each of these processes tends to be more like RBMT in implementation, though statistical probabilities can also play a part. Like SMT, one attraction of this approach to MT is the extent to which the computer can learn for itself how to do the translation.

By way of illustration, suppose we wanted to translate our earlier sentence (1). We first consult our database of already translated examples, and, assuming no exact match is found, we extract a set of similar matches. These might include (5)–(7) together with their translations as indicated.

(5) We apologize that there are no seats available.
Mae'n ddrwg gennym nad oes seddi ar gael.

(6) The late arrival of the train made us miss the start.
Collasom y cychwyn oherwydd fod y trên yn hwyr.

(7) The London train comes via Bristol.
Mae'r trên o Lundain yn dod trwy Bryste.

The task is now first to extract the relevant fragments from the examples: 'We apologize' from (5), 'the late arrival of the train' from (6), and 'the London train' from (7). Then we have to identify in the Welsh translations which bits of text correspond to the fragments. Finally, we need to "glue" these together to make the translation of the original input.

Each of these tasks has its difficulties. The "matching" task can be done in a relatively straightforward manner, if we use the simplest of character-matching algorithms (familiar from many other computer applications such as file comparison, spell checking and so on). A more linguistically sophisticated match is possible. For example, compare 'They apologize …' and 'We sincerely apologize …': the former is a closer match character for character, but the latter is more useful if we know how to accommodate the extraneous adverb.

The "alignment" stage is somewhat harder. It is so called because we have to try to align the English fragments that we are interested in with the corresponding Welsh fragments. Ideally, we need the computer to learn how to do this for itself. One way to do this is to compare further similar examples and extract the common elements. If we happen to have any linguistic resources such as dictionaries, this of course can be very helpful.

The "recombination" stage can also be difficult: the simple expedient of concatenating the fragments may result in translation errors due to "boundary friction", that is, the way the fragments need to fit together. An obvious example of this is agreement and mutation which might not be covered by the examples chosen. This can be a problem where the TL is considerably more complex than the SL, as may be the case for English and Welsh. Again, a way round this is to incorporate some elements of RBMT to "tidy up" the proposed output.

EBMT has been in development for about 15 years now, and is proving to be a viable approach

to certain types of translation task, again depending also on the language pair. Like with SMT, EBMT packages are readily available, and the only pre-requisite is large amounts of parallel text, though a machine-readable dictionary would also be helpful. In Section 4 **we will recommend development of a small English–Welsh EBMT system**.

2.2.5 SPOKEN-LANGUAGE TRANSLATION (SLT)

Of course the natural medium for language is speech rather than text, and perhaps the ultimate dream for MT is to have systems that can translate *spoken* language. This remains a tall order given the current state-of-the-art. Spoken-language translation (SLT) requires robust speech understanding (or speech recognition) and speech synthesis on top of accurate translation.

The first logical step in SLT is for the computer to process the acoustic signal it receives as input. The success of this process depends partly on the quality of the signal of course (quality of microphone, amount of background noise), but also on the nature of the spoken language, and the initial output required. We can distinguish "speech recognition", also known as "speech-to-text", where the goal is to transcribe the input, and "speech understanding" where the goal is to identify the speaker's intention. Interestingly, the latter is the less demanding task, with a bigger margin of error. Both tasks are made complicated by the extent that speech is not "spoken writing": natural everyday conversation contains hesitations, false starts, repetitions, incomplete sentences, and other disfluencies which the listener tends to filter out. These may be less prevalent in more formal spoken language.

The distinction between recognition and understanding underlies a distinction between two approaches to SLT. For SLT of more formal speech, it is plausible to consider concatenating speech-to-text, text-to-text MT and text-to-speech synthesis, the so-called "linear pipeline" approach. Much SLT research on the other hand has focussed on task-oriented cooperative dialogues (e.g. scheduling a meeting, consulting a travel agent) where the task is to understand what the speaker means, and to render this in the TL, without necessarily translating faithfully what was said. This is more like what a consecutive interpreter might do. The process relies more on matching what was "heard" as closely as possible to what can be expected, rather than recognizing individual words. SLT systems of this type are still very much experimental.

With the linear pipeline, although each of the processes is error-prone, some success has been reported with this simple scenario. Speech-to-text can be supported by a user for whose voice the system can be trained beforehand, and who can help the system by correcting the system's proposed transliterations. Translation can be made easier also by keeping texts simple, and by identifying the subject domain beforehand, enabling appropriate dictionaries to be loaded (e.g. tourism, going to the doctor, etc.). Of the three processes in the pipeline, speech synthesis is the most advanced: in an experiment by the present author,[9] it was shown that the comprehensibility of MT output is degraded by less than 1% when the text is output through speech synthesis, compared to 3% for speech recognition, and 12% when all three are combined.

Note that the conditions for all of these experiments involve a cooperative speaker speaking carefully and slowly, and waiting for the translation. At the current time there is little prospect of automating simultaneous interpretation. SLT for the time being means consecutive interpretation of simple dialogues. Whether this is of interest in the Welsh scenario is debatable. Nevertheless, the contributing technologies of speech recognition and speech synthesis for Welsh have other uses, notably

[9] H. Somers and Y. Sugita, 'Evaluating commercial spoken language translation software', *MT Summit IX: Proceedings of the Ninth Machine Translation Summit*, New Orleans, 2003, pp. 370–377.

for the disabled, and we can hope that work in these areas is ongoing, even if they are not immediately relevant to the present proposal for text-based MT between English and Welsh.

3 LT, WELSH, AND OTHER MINORITY LANGUAGES

In this section we wish to summarize the current situation regarding LT and Welsh, comparing it with a number of other languages, notably Irish, Catalan, Basque, and Galician.[10]

3.1 WELSH TRANSLATION NEEDS

The 2001 census indicated that 20.8% of the overall population of Wales "can speak" Welsh (i.e. about 582,000 people). The percentage varies hugely from region to region, with Welsh spoken by the majority of the population in four of the 24 unitary authorities (Gwynedd, Anglesey, Ceredigion, Carmarthen). Notwithstanding this, there are very significant absolute numbers of Welsh-speakers in the cities of south Wales where they form a reasonably small percentage of the population. The national figure has experienced a 2 percentage point increase since the 1991 census, reversing a long-running decline in numbers and percentages of Welsh speakers. The period between the two census dates also saw the enactment of the Welsh Language Act 1993 which established in law the equal status of English and Welsh. During this period, also, the National Assembly for Wales was established (1998).

While efforts continue to encourage and promote the use of Welsh, as a "minority" language it does not seem to be under serious threat of "extinction" (see Euromosaic report, 1996). From the point of view of LT, it falls into a group of "minority languages" which receive relatively less attention, mainly for economic reasons. This group typically covers (a) the European regional minority languages such as Basque, Breton, Catalan,[11] Romansch, (b) the languages of "small" countries like Bosnian, Croatian, and even Danish and Finnish, with relatively small numbers of speakers, and (c) languages which may have quite large numbers of speakers, but which are spoken in countries that have been considered not to be of economic significance particularly in the IT industry: included in this list are some of the world's most widely spoken languages (Hindi, Urdu, Benagli). Two other factors are relevant in the classification of Welsh: the relative difficulty of handling it in LT applications (low, because it uses the Latin alphabet with only trivial special features), and advantages accruing from its linguistic relation to other languages (few, because it is not closely related to any of the world's major languages). This mixed picture correctly predicts that low-tech LT support for Welsh is easy to envisage, but more sophisticated tools will require considerable investment. We will return to this below.

Part of our discussion of IT provision for Welsh will cover monolingual localization issues such as Welsh word processing, on-line language-related tools, speech processing and synthesis. But a major focus is the question of translation, and here we mean exclusively translation between English and Welsh.

A particular feature of the situation for Welsh, which it shares with several other "minority" languages in Europe is the fact that, apart from young children, effectively all Welsh speakers are functionally bilingual: while many Welsh speakers prefer to speak, read and write in Welsh (though this may depend to some extent on the subject matter), none are entirely unable to do so in English if necessary. This has an important ramification in terms of the perceived need to provide translation between English and Welsh.

[10] This section is based on a series of visits to a number of institutions and groups. I wish to acknowledge here the kind hospitality of Jeremy Evas, Mikel Forcada, Xavier Gomez Guinovart, Gwyn Jones, Donncha Ó Cróinín, Delyth Prys, Louisa Sadler, Kepa Sarasola, Andy Way and their many colleagues.

[11] The level of resourcing for these languages varies greatly: compared to Romansch, say, Catalan is rather well off.

Bearing this point in mind, translation *from English into Welsh* is provided for the convenience of Welsh speakers by, amongst many others, 260 organisations with statutory language schemes, e.g. public bodies, local councils, the courts, health authorities and so on. Translation is of general announcements, increasingly nowadays on web sites as well as leaflets and mailshots. The Official Record of the National Assembly, published in full bilingual version five days after the debates, meetings and so on that it records, and is "fully bilingual". All proceedings are translated into either English or Welsh (cf. the point made in the previous paragraph); verbatim transcripts of plenary sessions are published within 24 hours, with translation only of Welsh into English. Translations need to be of high quality. Although the Record of Proceedings is translated by a commercial company external to the the Assembly, the NAW itself employs 45 translators. Local authorities also have small in-house translation departments, Gwynedd for example has five or six translators. Independent translators are widely used.[12]

Individuals writing to public bodies in Welsh can expect a reply in Welsh, though this may not involve translation when there are Welsh-speakers in the organization who can handle the correspondence directly in Welsh, or where standard reply letters have already been drafted.

Translation and interpeting services *into English* are provided in the Assembly and by some local authorities for the benefit of non-Welsh speakers in situations where Welsh is being used, for example in emails and other less formal texts. Rough (gist) translation in this direction seems to be acceptable, and MT could certainly play a role here. Simultaneous interpretation of Welsh into English (but not vice versa) is also provided in the Assembly and in some meetings. Of interest possibly is the shift in perceptions, perhaps similar to that in Belgium over the last 50 years, where speakers of the formerly dominant language have to accommodate the use of the other language rather than vice versa. The distinction between interpreting and translation services needs to be borne in mind: the latter implies that written documents are originating in Welsh, which was probably not the case even 10 years ago.

3.2 LT AND WELSH: CURRENT PROVISION

In this section we will briefly review the current provision of LT for Welsh. It is convenient here to adopt the headings used in the present author's review of LT for minority languages.[13]

3.2.1 WORD PROCESSING

Welsh uses the same alphabet as English, with the exception of accented characters 'ŵ' and 'ŷ', which are catered for in "Latin extended-A" character set available with most Windows-based text-input programs. Purely mechanical aspects of word-processing such as justification and fonts do not require any adaptation for Welsh. Other aspects of word-processing reflect linguistic differences. There are alphabetization differences between English and Welsh involving digraphs 'ch', 'dd', 'ff' etc. Hyphenation rules for English do not apply to Welsh. Automatic typographical effects (e.g. superscripting of 'th' in ordinal numbers) are language-specific. 'Welsh' is included in the list of languages available in the Tools menu in Microsoft Word XP and 2003. This enables the use of the Microsoft Welsh language proofing tools. This will work independently, for Office XP, or together with the soon to be released free Welsh Language Interface Packs (LIPs) which will localise Microsoft

[12] See www.welshtranslators.org.uk.

[13] Somers, H. 'Machine Translation and minority languages'. *Translating and the Computer 19: Papers from the Aslib conference* (London, November 1997).

Windows XP and Office 2003. These will be available, thanks to a partnership between the Welsh Language Board and Microsoft announced in January 2004.[14]

3.2.2 SPELL CHECKERS, DICTIONARIES AND THESAURI

Spell checkers for languages like English operate by a simple algorithm involving an exhaustive list of possible words: any word not in this list is assumed to be misspelled (unless it begins with a capital letter), and a suggested correct spelling is identified by permuting the letters typed, and allowing for some insertions and deletions. For morphologically more complex languages, where for example the list of possible words is theoretically infinite, spell checking must include some language-dependent linguistic processing. ***Cy****sill*™, a spell checker and hyphenator for Welsh, developed by Canolfan Bedwyr,[15] has received wide acclaim. As well as a spell-checker, this software also works as a grammar checker and includes hints and explanations in both English and Welsh which make it suitable as a learner's tool as well. It uses a linguistic rule formalism and pattern-matching templates using regular expressions. Implemented in C++, it is highly portable.

Dictionaries, in the normal sense of the word, are much more than word-lists: as well as distinguishing different word senses, they will usually offer some grammatical information. In one sense they are also something less than a word-list, since they usually do not list explicitly all the inflected or derived forms of the words. It is important to distinguish monolingual, bilingual and multilingual dictionaries, and specialist terminologies. Structure terminologies are called "thesauri", not to be confused with the popular notion of a "thesaurus" as a list of synonyms and related words.

Canolfan Bedwyr published *CysGair*, an electronic bidirectional Welsh–English dictionary, in 1995, available on CD since 2002. It covers approximately 35,000 words, including idioms and proverbs, and has been compiled in part from a dictionary publisher's typesetting tapes. They also adapted the 1996 Welsh Academy English–Welsh dictionary to electronic format for publication as a web-based dictionary. This is the "BBC dictionary" referred to in footnote 3. Several terminological dictionaries have been developed, some of them in electronic form. Electronic and web-based dictionaries generally need to be interfaced with software that can handle mutations, regular and irregular plural forms, and verb inflections. Canolfan Bedwyr's *Cysgliad* package includes a dictionary and terminology package *Cysgeir*, said to be the "next generation" of *CysGair*.

3.2.3 CORPUS-BASED RESOURCES AND TOOLS

Much language technology work nowadays focuses on huge collections of text ("corpora") and tools that can be developed from them. Monolingual corpora consisting of examples of "real-life" use of language are used for example by lexicologists to check usage of words and phrases, and by language teachers for genuine illustrations of language use. An important tool for this purpose is the "tagger", a piece of software that analyses text and assigns part-of-speech (POS) "tags" to the words. Taggers can be rule-based, like traditional linguistic grammars, or they can be statistical, "learning" likely POS sequences from training data. Bilingual or parallel corpora are essential for the development of SMT and example-based MT systems, as described above.

---

[14] See http://www.welsh-language-board.org.uk/en/cynnwys.php?cID=6&pID=6&nID=220

[15] Canolfan Bedwyr was established in 1996 as a support service at the University of Wales, Bangor. Development of terminology and language technology for Welsh is one of its activities.

Canolfan Bedwyr are developing a tagger, but work is still at an early stage. In 1991 they assembled the CEG (Cronfa Electroneg o Gymraeg) corpus of just over 1m words, based on 500 samples of approximately 2,000 words each, selected from a representative range of text types to illustrate modern Welsh prose writing.[16] They are also currently assembling an informal parallel corpus.

The BIML (British Indigenous Minority Languages) project[17] at Lancaster University was funded by the EPSRC and included in its aims the collection of a small (80,000-word) spoken corpus of Welsh, and the development of a Welsh tagger. The website[18] provides links to downloadable resources associated with the tagger, including the tagged corpus itself. Their report[19] on existing resources also gives a number of suggestions and URLs for mono- and bilingual websites which are an obvious source of text data.

Some work on Welsh is being undertaken in a more traditional computational linguistic framework at the University of Essex. This is within the PARGRAM (Parallel Grammar) framework, and is part of a 3-year research project (started April 2004) funded by the ESRC. The aims of the project[20] are somewhat limited, since the main interest is applying the theories developed within the PARGRAM framework to a verb-initial language. The Essex team expect to concentrate on the major grammatical constructions, with a limited lexical coverage, based on a small reference corpus.

3.2.4 MT SYSTEMS

A search of the WWW for MT systems for Welsh offers numerous links to the *InterTran* system, from London-based Translation Experts Ltd.[21] Their website[22] offers free translation, but repeated attempts to try the system were met with "over quota" warnings, and an invitation to subscribe. It is highly unlikely that the software can offer anything more than word-for-word substitution for such a range of languages.

Some other research on Welsh MT is reported. As mentioned above, John Phillips of Yamaguchi University in Japan, presented a paper[23] at the 2001 PACLING conference (Kitakyushu, Japan) on an SMT system trained on the Bible. The system benefits from an additional bilingual dictionary, and morphological analysis. He exemplifies its translation quality with a long example from a Welsh novel, the first sentence of which is shown in (8). The quality of the translation is remarkably good considering that the system was trained on a completely different text type.

---

16 Source: http://www.bangor.ac.uk/ar/cb/ceg/ceg_eng.html.

17 See http://www.ling.lancs.ac.uk/biml/.

18 http://www.ling.lancs.ac.uk/biml/bimls3corpus.htm

19 http://www.ling.lancs.ac.uk/biml/bimls3reports1.htm

20 See http://privatewww.essex.ac.uk/~louisa/esrcproj/, http://www.dcs.kcl.ac.uk/staff/mary/pargram/.

21 See http://www.tranexp.com/InterTran/FreeTranslation.html. TranExp offers tools which translate between 1482 language pairs, Welsh among them.

22 http://www.tranexp.com:2000/InterTran/

23 See footnote 8.

(8) Gorweddai hi ar lan yr afon, ei gwallt ar led o'i chwmpas, ei dau lygad ar gau.
She would lay upon the river, her hair, spreading around her her eye both close.

Llio Bryn Humphries is also reportedly developing an SMT system[24] as part of a part-time MPhil in the Computer Science Department, UW Bangor. A brief telephone discussion indicated that she has no concrete results yet, but expects to finish her work in about a year (Summer 2005).

3.3 MINORITY LANGUAGES ELSEWHERE IN EUROPE

In this section, we report on a round of visits to five European locations where there is a language situation possibly in some way comparable to that of Welsh. In the course of this project we visited Ireland, Catalonia, Galicia and the Basque Country. Coincidentally we were also able to visit Malta during this period.

3.3.1 IRISH

The situation in Ireland is superficially similar to that in Wales: the majority of the population is English-speaking, and there are effectively no Irish monolingual speakers. Irish has the status of first official language, and – unlike Welsh – is recognized within the EU as a "treaty language". However, Irish is less widely spoken than Welsh, despite its more official status. The 1996 census discovered that 71,000 adults said they spoke Irish every day. Of these, almost 21,000 lived in the Gaeltacht, but realistically the number of people using Irish as their main language in the mid-1990s was probably between 10,000 and 21,000.[25] Interestingly, according to the 1991 UK Census, there were 79,012 people in *Northern* Ireland able to speak, read and write Irish, and 45,338 claiming to speak the language. Around 10% of the total population of Northern Ireland had some knowledge of Irish.[26]

The Dáil and Seanad proceedings are all translated into Irish, but this may take up to 3 years. Of the 166 deputies only 3 or 4 might make a speech in Irish, and prior notice would have to be given: interpreters are not routinely provided. In the courts, Irish-speaking judges are available on request, but in the health service it might be hard to find doctors and nurses who speak Irish. All children learn Irish at school. Irish is used as the language of instruction in the Gaeltacht (pop. 83,000), and in Gaelscoileanna (Irish-speaking schools). The state broadcaster RTE operates an Irish-language channel, TG4 showing many dubbed and subtitled programmes, especially for children.

The LT provision for Irish is considerably inferior to that for Welsh. There is a spell-checker, developed jointly by the ITE (Institiúid Teangeolaíochta Éireann)[27] and Trinity College, and distributed by Microsoft. It does not however include a morphology or grammar checker. Unlike Welsh, as yet there is no widely available Irish-language interface for Windows, though an earlier Mac o/s, and Claris works, were localized for Ireland. The future looks rosier however, as Irish is included in Microsoft's LIP programme.

---

[24] http://www.informatics.bangor.ac.uk/~llio/progress-report-ii-e.htm

[25] Source: http://homepage.ntlworld.com/r.a.mccartney/baile_nua/numbers.html

[26] Source: http://www1.fa.knaw.nl/mercator/regionale_dossiers/regional_dossier_irish_in_northernireland.htm

[27] Institiúid Teangeolaíochta Éireann (The Linguistics Institute of Ireland) existed between 1973 and February 2004, funded by the government, to provide research and advice on all aspects of language.

Online resources are fairly limited. Parliamentary proceedings are available in machine-readable form, though not on the Web. The laws of Ireland (the Acts of the Oireachtas) are available on-line[28] and thus form a considerable parallel corpus. There are some Irish texts in connection with the EU, and some universities and organizations have Irish pages. The ITE has collected the 30m-word National Corpus of Irish,[29] a small subset of which (200,000 words) is POS-tagged. Work is in progress on a lemmatizer/tokenizer.

There is LT expertise in Ireland, at Dublin City University (DCU), Limerick and Trinity, and some lexicographic work is being done at the Royal Irish Academy and at Coleraine. As far as MT is concerned, the main thrust is basic research in the School of Computing, DCU collaborating with translation specialists in the School of Applied Language and Intercultural Studies. Their main interest is in EBMT and CAT tools.

3.3.2 CATALAN

Catalan is spoken in the autonomous communities of Catalonia, Valencia and the Balearics in Spain, in Andorra, in bordering regions of France, and in the town of Alghero in Sardinia. In the Spanish regions it is a co-official language, or *llengua propia* ('own language'). Of a population of 11m, 7–8m know Catalan, and 6m identify themselves as Catalan speakers. For various reasons, Valencian is sometimes recognized as a different language. About 15% of the schools in Valencia are Catalan-speaking, and Spanish is taught as a second language. In Catalonia, this figure is higher than 90%.

Catalan has official status in that citizens can use the language to communicate with authorities In general there is a policy of bilingualism. Spanish is more prevalent in Valencia, but in Catalonia, the *llengua propia* is more prevalent, sometimes even exclusive. Immigrants to the region will typically learn Catalan rather than Spanish.

A major difference between Catalan-speaking regions and Wales is of course the fact that Spanish and Catalan are relatively closely related languages. All Catalan speakers will understand Spanish, though not necessarily vice versa. The view that Catalan is a dialect of Spanish is quite wrong: historically in fact Catalan is nearer to French and Italian, being derived from Vulgar Latin while Spanish derives from Classical Latin. There are consequently large lexical differences, although the syntax is very similar. From the point of view of translation, particularly MT, this means that an approach which is essentially word-for-word replacement can deliver quality levels around 80%.

LT provision for Catalan is quite good. A localized Microsoft o/s is available as a patch, and a lot of localization work has been done for open source software. Catalan, again, forms part of the Microsoft LIP project mentioned above. There is a decent spell-checker for Catalan, but not for Valencian. There is no grammar checker available (even the best grammar checker for Spanish is not very good). Parallel text (mostly Spanish and Catalan, rarely with English as well) is abundantly available on the Web. The bilingual *El Periódico de Catalunya* is published daily online (formerly for free, now by subscription). There are several MT systems translating between Catalan and Spanish. Probably the best is *Comprendium.es* which also can translate Catalan–English. Its development was funded with approx. €2m from the Catalan Generalitat (government). Another system, *Automatic Trans*, was first developed to translate *El Periódico de Catalunya*, but is now available commercially. Autotrad have two systems, *Ara*, which sells for just €50, and *SALT*, translating Spanish–Valencian using the same engine as *Ara*, paid for by the

[28] http://www.achtanna.ie/

[29] http://www.ite.ie/corpus/

Valencia government, and free to users. Quality is good, but it is very slow. *InterNostrum* is perhaps the biggest system. Developed over 20 person-years since 1999, costing €200,000 thanks to various funders, it has 20,000 words in the dictionary and translates very fast at around 5,000 words/second.

Of all the minority languages, Catalan is in the best shape from the point of view of LT. This is probably because the language is important – around 8.5m speakers (that's more than Danish) based in Spain's financial capital – but also because the technical difficulty of translation between Catalan and Spanish is much reduced. The *InterNostrum* team for example were able to develop a Portuguese version of their program in just 4 months.

3.3.3 BASQUE

The Basque language, unrelated to any other European languages, is spoken in Northern Spain and the adjoining area of Southwest France. Seven or eight distinct dialects have been identified, though a single standard has emerged at least for written Basque. According to one source,[30] about 24% of the overall population of the Euskal Herria (the Basque Country) speaks Basque, with higher concentrations in rural areas (up to 65% in Navarre), down to much smaller numbers of speakers in Bilbao, and even fewer in French cities such as Bayonne and Biarritz. On the Spanish side of the border, Basque has official status in a way similar to Catalan; it has no status as such in France.

Like Welsh, Basque is used as a working language, and is the language of instruction in some schools. Again, paralleling the Welsh situation, all Basque speakers can also function in Spanish, so provision of Basque is seen as meeting a kind of human right. Educated Basque speakers often have English ability as well. Most translation is from Spanish into Basque for official purposes. In general, interpreting services are not much called upon.

LT provision for Basque is quite good, thanks in part to the Basque Government's sponsorship (they initially paid Microsoft to develop tools for the language, but now Basque, like many languages mentioned above, is part of Microsoft's LIP scheme). A spell checker has been available for more than 10 years. This was developed at IXA in the Euskal Herriko Unibertsitatea (University of the Basque Country) in Donostia/San Sebastian,[31] and has contributed to a degree of standardization between the dialects. Lexical tools such as dictionaries and thesauri have had to incorporate sophisticated morphological components, since Basque (even more than Welsh) is highly inflected.

There are a number of groups involved in research and development on Basque LT, the most significant of which is the IXA group already mentioned. Active since 1988, the group now numbers around 50 people. A recent initiative, Hizking21,[32] promotes collaboration between university-based research groups and a number of private companies.

Research and development has followed a strictly phased stratificational pattern, first laying down the scientific foundations, then developing basic tools before integrating these into "products", initially of modest complexity, then more sophisticated.[33] An MT system counts among the last of these, so, despite its long history, the IXA group has only recently started work on MT. Initially they are working

[30] http://www.nabarralde.com/anabarra/anabarra_bsa.html

[31] See http://ixa.si.ehu.es/Ixa. The name is not an acronym, just the Basque name for the letter 'X'.

[32] See http://www.hizking21.org/hizking21/publikoa/Sarrera?lang=ing

[33] See http://ixa.si.ehu.es/Ixa/Aurkezpena

on Spanish–Basque, and will later work on English–Basque. Interestingly, there are no plans to develop MT with Basque as the SL. Like their general philosophy, the approach to MT is strictly bottom-up: using tools already developed, notably morphological analysers and lexical tools, a word-for-word approach was first developed which has now been refined so that noun-phrases are handled. Work is currently focused on verb groups, and simple sentences, while the further complexities of morphological structure await attention. As well as the traditional rule-based approach, there is some interest in EBMT, which they see as being an extension of proposed work on TMs and other translator's tools.

3.3.4 GALICIA

Galician is spoken in the province of Galicia and some neighbouring regions in Northwest Spain. Linguistically it is closer to Portuguese than Spanish, and, like other regional languages in Spain, has official status in its own province. Of a population of 2.7m in 1991, it is claimed[34] that 94% understand Galician, and 88% speak it, though a somewhat smaller number (55%) consider it their primary language. More significant though is the fact that usage is much lower among the middle class and young people,[35] and despite these high numbers, Galician is in a state of diglossic bilingualism with Spanish dominating. The low social status of Galician is only recently being addressed with its use in education, administration and the media.

LT tools and resources for Galician are quite modest. Galician forms part of Microsoft's LIP programme. IT includes a spell checker, developed by Imaxín Software,[36] and funded by the Galician Government (Xunta de Galicia). Little else is available, though three MT systems exist, all translating Spanish–Galician. One is the *ES-GA* system, developed by the Xunta's research centre CRPIH (Centro Ramón Piñeiro para a Investigación en Humanidades) in Santiago de Compostela. The system is based on the well-established rule-based *METAL* system, though running on a workstation rather than a PC limits its commercial viability. Another system, *TraduZa-g*, has been developed with private funding by the Santiago-based company Dimensiona,[37] and drives the free on-line translation service offered by the media server Mundo-R.[38]

Development of Galician LT has been hindered by the relatively low enthusiasm for the Galician language among the influential middle class and young speakers in the province (compared, say, to Catalan and Basque, or Welsh for that matter). On the research front, the Xunta-based CRPIH in Santiago is well funded and influential in decisions about research topics, directions and results; research is also undertaken at universities in Vigo and A Coruña, who also collaborate with Portuguese researchers at the Universidade do Minho in Braga, and Porto University.

3.3.5 SOME OTHER CASES

We have seen a contrasting range of cases in the preceding paragraphs, and have focused on languages where there is at least some significant LT provision. One other case worth mentioning is Malta.

---

[34] Source: Xulio Sousa, Instituto da Lingua Galega, Universidade de Santiago de Compostela, http://www.usc.es/~ilgas/galician.html

[35] *Lingua inicial e competencia lingüística en Galicia*, Real Academia Galega, A Coruña, 1994; cited in F. Lema-Alvarellos, *MT for Minority languages: the case of Spanish-Galician*, MSc dissertation, Department of Language and Linguistics, UMIST, Manchester, 2003.

[36] http://www.imaxin.com/

[37] See http://www.dimensiona.com/traduza/default.htm

[38] See http://traductor.mundo-r.com/

Coincidentally during the preparation of this report, the author attended an MT conference in Malta just days before that country joined the EU. This accession has, among other things, meant that there will be significant interest and investment in Maltese, which is now an official EU language. This is of course slightly anomalous: Malta's population is only 400,000 of which some 300,000 are Maltese speakers (the majority, but not all, also English speakers).[39] This is about double the number of Irish speakers, and of course 200,000 *less* than Welsh, which has no EU status. There is a small group of NLP specialists in the University of Malta who will now presumably receive great encouragement to develop LT tools for Maltese.

In contrast, let us consider finally the other Celtic languages, for most of which what support in terms of LT they have is mostly the work of enthusiasts, since they lack official status in the country where they are spoken.

The Celtic languages divide into two groups, the Brythonic or "P-Celtic" languages, Welsh, Breton and Cornish, and the Goidelic or "Q-Celtic" languages Irish, Scots Gaelic and Manx. The two groups are named according to whether or not they underwent a sound-change from underlying 'qu' to 'c' or 'p' (compare Irish *ceann* and Welsh *pen* 'head').

Since language questions are not included on French censuses, the number of Breton speakers can only be estimated. One website[40] suggests as many as 300,000 speakers and a further 150,000 who can understand Breton. Despite some public support for the language, it has no official status in France and, more worryingly, numbers of speakers are in sharp decline, with fewer than 2,000 speakers under the age of 25, leading to the conclusion that "it is likely Breton will die out in the next half century".[41] Nevertheless, there appears to be a degree of computer support for Breton: the Korvigelloù An Drouizig webpages[42] detail a spell-checker and grammar-checker, some localized software, and some language-related computer games.

Cornish became a "dead" language in the 18th century but has been revived by enthusiasts so that there are now about 3,500 speakers, including some native speakers.[43] As yet it has no official status in the UK. A Cornish spell-checker, *KernSpell*,[44] has been developed (by the Dublin-based Everson Typography), but as yet only for Mac users.

According to the 2001 census, Scots Gaelic has some 58,650 speakers (an 11% decrease compared to 1991) mostly living in the Highlands and islands. Gaelic is taught in some schools, including some in which it is the medium of instruction, and is used by the local council in the Western Isles, Comhairle nan Eilean.[45] Official recognition in the form of a Gaelic Bill is being considered by the Scottish

---

[39] Source: http://www.intersolinc.com/newsletters/newsletter_37.htm

[40] http://ww2.eblul.org:8080/eblul/Public/member_state_committ/french_committee/france2/view

[41] Source: http://www.sciencedaily.com/encyclopedia/list_of_endangered_languages

[42] http://www.drouizig.org/drouizig/

[43] Source: http://www.sciencedaily.com/encyclopedia/cornish_language

[44] http://www.evertype.com/software/mackernspell/. Everson have also developed spell checkers for Gaelic and Manx (see footnotes 46 and 49), though presumably none of these products has the sophisticated grammar/mutation checking found in ***Cysill***™.

[45] http://www.sciencedaily.com/encyclopedia/scottish_gaelic_language

Parliament. Very little Gaelic software exists, though there is a spell-checker, *GaidhealSpell*,[46] also developed by Everson, again only available for Mac users.

By the time the last Manx native speaker died in 1974, there was already a revivalist movement. The Manx government is supportive, with the language offered as an option in all junior and secondary schools on the island, and in 2001 Manx has been available as the language of instruction to a small group of infants sponsored by the parent group Sheshaght ny Parantyn.[47] Further government support is evidenced by the use of Manx in the Tynwald, with new laws being read out by Yn Lhaihder ('the Reader') in both Manx and English.[48] *ManxSpell* is a Manx spell-checker for Mac users has been developed for Mac users by Everson.[49]

3.3.6 CONCLUSION

LT support for minority languages seems to vary, not always in proportion to actual needs, or even official status. In fact, the situation for Welsh seems to be the best of the Celtic languages. From the point of view of MT, Catalan and to a lesser extent Galician are favoured by being closely related to Spanish, which is the TL for which translation is needed. It is perhaps surprising that there is no Irish–English MT system, given the official status of Irish and the fact that there is NLP expertise and interest in the country. The situation of Basque is probably the most salutary: an early decision to develop tools in a stratified incremental manner has meant that there is a good infrastructure in place, but not much software to show non-specialists, perhaps to their dismay. This is true to a lesser extent for Welsh, with the highly visible and much acclaimed grammar-checking software underpinned by solid NLP research. **The time is right to develop MT to a quality comparable with more established systems, and the expertise and software support is already in place.**

[46] http://www.evertype.com/software/macgaidhealspell/

[47] http://www.gaelg.iofm.net/SCHOOL/info.html

[48] http://www.sciencedaily.com/encyclopedia/manx_language

[49] http://www.evertype.com/software/macmanxspell/

4 DEVELOPING MT FOR WELSH

In this section we wish to make concrete recommendations for the development of MT software for Welsh. The proposed strategy will be for three types of system to be developed in parallel with a view in the medium-term to integrating all three in a single multi-engine MT system.

More specifically, we would like to **recommend the funding of a tripartite approach to developing MT** based on the three currently favoured technological approaches (rule-based, statistics-based and example-based), **with a shared common infrastructure**, and **with a view to integration in a multi-engine system in a second phase,** which would also include continuing refinement of the existing systems.

Our recommendation is for a first phase **two-year project** with a budget split between **three branches**, with a **project director** on a consultancy basis. In the following sections, we name several ***possible* collaborators** identified, on the understanding that any such work would be **subject to the normal tendering process**, and that other suitable candidates might be thus identified. In order to ensure that all possible tenders are received, it is recommended that information about the research be **announced on appropriate websites and bulletin boards**, notably mtlist.[50]

4.1 SMT SYSTEM

To develop an SMT system, a *training set* of translated sentence pairs is required: at least 100,000 words in each language, preferably 1m or more. This data must be converted to plain text format, and must be sentence-tokenized, and sentence-aligned. Each of these tasks involves some human effort, either running and supervising programs, or checking (a random selection of) the results for accuracy, and the tasks may or may not require someone with knowledge of Welsh, depending on the language direction chosen (see below).

In addition, an evaluation data set is required, consisting of a set of multiple reference human translations for use with Bleu or other evaluation metrics. A corpus of between 250–1000 sentences translated by 3 or 4 human translators is recommended. Depending on the source of the data, one set of translations may already be available. In addition, around 300–500 translated sentence pairs must be word-aligned, a task which nevertheless requires a Welsh-speaker's human judgment.

The SMT process is essentially reversible, so the same data can be used to build both a Welsh–English and an English–Welsh system.[51] The decision of which system to build in the first place rests on two factors:

(a) What is the expected quality of translations produced by an SMT system, and to what use translations of such quality could be put.

(b) What aspects of the procedure require a Welsh speaker.

Evidence so far suggests that SMT systems can produce understandable but relatively low quality translations, often with minor grammatical anomalies that the statistical process (with no linguistic "knowledge" incorporated into it) cannot pick up (but which can be corrected by other processes, for

[50] http://www.eamt.org/mt-list.html

[51] Note that this is not meant to imply that the translations output will be mutually identical.

example in a multi-engine approach). For this reason, **we would recommend in the first instance that contractants be invited to demonstrate the feasibility of their approach, building a small Welsh–English SMT system for gist translation of Welsh documents**. This imposes a burden of either locating or producing suitable parallel texts – 1m words or more is needed. Suitable texts that might be relatively easy to find would be informative web pages of Welsh institutions. Help would undoubtedly be required from a Welsh speaker to identify such pages. We assume **this task could be achieved within 9 to 12 months**. At the end of this period, they will be asked **to evaluate the shortcomings of the system, and if possible suggest improvements**.

Assuming that the results of the first experiment were satisfactory, they could then be invited **to use the same methodology to develop an English–Welsh SMT system**. It is appreciated that the need for English–Welsh translation is for higher quality of output for a different type of input, speeches and more formal documents, for example. The development of an English–Welsh SMT system would be seen as a contribution to be integrated into a multi-engine system, to be developed in a second phase.

4.2 EBMT

To develop an EBMT system, an *example set* of translated sentence pairs is required. This requirement is similar to that for the SMT system, but the volume of material required is more modest: a minimum of 1,000 sentence pairs, though the system will work correspondingly better with more data. For EBMT there is more of a premium on homogeneity of corpus data, but effectively the same data (e.g. university and local council web-pages) will be acceptable. The alignment work is similar to that needed for the SMT system, but not identical, not least because we will be working initially on the English–Welsh translation direction.

After the sentence alignment work, a second phase of sub-sentential alignment is required, followed by adaptation of the recombination algorithm. Automatic evaluation can use the same reference translations as have been provided for the SMT experiment, though some human evaluation is also indicated.

The EBMT process is essentially language-independent, so again we have a choice of whether initially to build a Welsh–English or an English–Welsh system. Like SMT, EBMT systems can produce understandable but relatively low quality translations..

4.3 RBMT

We propose the development of a Welsh–English RBMT system. Unlike the other two systems, there is a considerable human expert effort required in writing linguistic rules, besides the programming and testing. Furthermore, the task cannot be neatly divided into subtasks which can be individually budgeted. Already existing tools can be adapted and extended for the task of MT, such as the rule-based ***Cy****Sill*™ spelling and grammar checker, the lexicon, and so on. It should also be acknowledged that this development cannot be achieved in the short time scales needed for SMT and EBMT.

4.4 PROJECT MANAGEMENT

Because it is spread over three centres, **the overall project needs an amount of coordination and management**. This is particularly needed at the start, and towards the end when a follow-up project will be proposed, involving integration of the three systems into a single multi-engined MT system. Report writing will also need to be coordinated. Some cross-fertilization should certainly be encouraged, and **we would propose holding a 2-day Workshop** at the start of the project involving all contractors as well as the Welsh Language Board, at one of the centres. **Some funds for travel** can also be earmarked, and for presenting the project at relevant conferences, for example the 5th Celtic Linguistics Conference,

scheduled for September 2005.[52] An end-of-project Workshop will be held, possibly open to outside participants (perhaps in connection with a relevant national or international CL or MT conference, if in a suitable location).

4.5 MEDIUM- AND LONG-TERM PROSPECTS

The present report makes recommendations for a two-year programme to develop in parallel three types of MT system for reasonable quality Welsh–English translation, with lower quality English–Welsh translation provided by two of the engines. On completion of this project, an immediate goal is to integrate the different engines into a single system providing a quality of translation as good as that provided by the best state-of-the-art MT systems for other languages. Another issue to consider is "embedding", such as making MT available on the WWW, assuring compatibility with commonly-used software, for example Word, Excel, Oracle and so on (as well as non-Microsoft products). This is a task for "software engineering" rather than LT as such.

Other goals that might be envisaged in parallel with this activity (though not included in the present recommendation) include improving existing tools and providing further appropriate **tools for translators** working with Welsh, such as electronic dictionaries, proper-name lists, spelling and grammar checkers, sophisticated TM systems (linking up especially with work on EBMT), but also speech-to-text dictation and speech synthesis.

The last of these involves Welsh **speech recognition**, which can be seen as more of a long-term goal, along with Welsh **text-to-speech synthesis** (e.g. for blind people), as well as synthesis from abstract representations as might be found in SLT, as described in section 2.2.5 above.

We can summarize these future developments as follows:

- Integration of separate MT engines as a single system
- Integration (embedding) of MT in other applications
- Development of tools for translators
- Development of speech recognition and speech synthesis, eventually SLT

A much-used aphorism is Confucius's statement that a journey of a thousand miles begins with one small step. In fact, the steps proposed here are not so small, but would represent a significant leap towards bringing Welsh in line with the world's major languages in terms of LT provision.

4.6 SUMMARY OF RECOMMENDATIONS

1. Coordinated development of three-engine MT system:
    a. SMT, initially Welsh–English, then English–Welsh
    b. EBMT, initially English–Welsh, then Welsh–English
    c. RBMT concentrating on Welsh–English

[52] See http://privatewww.essex.ac.uk/~louisa/celtic/clc4.html.

2. Management role to coordinate the three approaches, to ensure compatibility and possible integration, and to report to sponsors.

3. Two workshops, at project start-up (internal only) and at the end of the project, open also to other relevant researchers, perhaps as part of an existing conference series.

5 BIBLIOGRAPHY

The following list includes some appropriate further reading matter, including items not referred to in the text.

Lynne Bowker. *Computer-Aided Translation Technology: A Practical Introduction*, Ottawa (2002): University of Ottawa Press.

Michael Carl and Andy Way (eds) *Recent Advances in Example-Based Machine Translation*, Dordrecht (2003): Kluwer.

John Hutchins. 'Machine Translation: General Overview' in Ruslan Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*, Oxford (2003): Oxford University Press, 501–511.

Kevin Knight. 'Automating Knowledge Acquisition for Machine Translation', *AI Magazine* 18 (1997).

Harold Somers (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam (2003): Benjamins.

Arturo Trujillo. *Translation Engines: Techniques for Machine Translation*, London (1999): Springer.

Stephan Vogel, Franz Josef Och, Christoph Tillmann, Sonja Nießen, Hassan Sawaf and Hermann Ney. 'Statistical Methods for Machine Translation', in Wolfgang Wahlster (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin (2000): Springer Verlag, 377–393.

Workshop on Language Resources for European Minority Languages Workshop at the First International Conference on Language Resources and Evaluation (1998), Granada, Spain.

Workshop on Developing Language Resources for Minority Languages: Re-useability and Strategic Priorities, Workshop at the Second International Conference on Language Resources and Evaluation (LREC 2000) Athens, Greece.

Workshop on NLP of Minority Languages and Small Languages, Workshop at TALN 2003, Batz-sur-Mer, France.