Automated English subtitling of Welsh TV Programmes

Llio Humphreys
Testun Cyf.
Norfolk House
57-59 Charles Street
Cardiff
Wales, UK
CF10 2GD
llio@testun.co.uk

Abstract: This paper discusses a government-funded Knowledge Transfer Partnership (KTP) project between Testun and Canolfan Bedwyr. The aim of the project is to develop an automated system for automating English subtitles for Welsh TV programmes, integrating a Welsh speech recognizer and a Welsh-English translation module within a subtitling environment. The paper discusses the motivation and mechanism, challenges, project plan and scope of this ambitious project as well as references to relevant prior research.
Keywords: subtitling, speech recognition, machine translation

1 Introduction

1.1 Motivation

Testun is a translation, subtitling and teletext company based in Cardiff, Wales. Testun's subtitling department edits prepared subtitles, and provides live and late subtitles, for programmes on S4C's three Welsh TV channels.

S4C has a target to subtitle 100% of programmes by the end of 2009, where copyright permits, so that more non-Welsh speakers and hard-of-hearing people can enjoy more Welsh programmes. The number of programmes with live sections is also increasing. There is an increase in demand for Testun's work. The problem is that translation, subtitling and teletext is labour-intensive.

Testun envisages that an automated translation and subtitling software solution would help Testun provide a more efficient subtitling service.

1.2 Mechanism

The Knowledge Transfer Partnership (KTP) Scheme helps UK businesses increase their competitiveness and productivity by accessing knowledge, technology and skills within research institutes. This project's research institute, Canolfan Bedwyr's Language Technology Unit, develops language resources for the Welsh language, the Celtic languages, and for multilingual situations in general. The Unit is responsible for standardising terminology, dictionaries, a Welsh language spellchecker and grammarchecker, computer-based learning and speech technology.

1.3 Challenges

The project involves challenges in intellectual property, security and software integration. Most of Testun's subtitling work takes place in S4C's offices using S4C equipment chosen and purchased by S4C. The automated subtitling software must be compatible with existing systems. Comprehensive consideration of these issues is beyond the scope of this paper.

The critical functional challenge is making best use of advanced, but brittle technologies, to produce subtitles comparable in quality to human-generated subtitles.

Speech recognition is made difficult by:-

- differences in voice quality, between men and women, young and old
- differences in accents and pronunciation
- mumbling
- interference: music, coughing etc.
- unknown words

Challenges for translation are:-

- good use of bilingual corpora - accurately aligning words or phrases where two languages express things

differently, using different parts of speech
- processing grammatically incorrect spoken text
- segmenting sentences with disfluencies
- dealing with unknown words
- producing subtitles that conform to industry-standard regulations

The challenge of providing NLP (Natural Language Processing) solutions involving less resourced languages has necessitated this project. The lack of available NLP resources does not mean that lower standards will be acceptable to S4C or its audience, who are used to high-quality human-generated subtitles. S4C's contracts for live and late subtitling (where pre-recorded programmes are received for subtitling shortly before transmission) expressly require a high standard of translation, and a spelling and grammar accuracy rate of 95%.

One solution is to view the problem of producing subtitles as much as information retrieval as language generation. For each sentence spoken, a search is conducted for the most similar Welsh sentence from a domain-specific parallel corpus, from which the English translation provides the subtitle output. Another solution is to limit the language domain, since both speech recognition and translation software perform best on sublanguages. Uncertainty from the speech technology side can also be reduced by loading a script for the TV programme. The speech module then aligns the spoken text to the script, to ensure appropriate subtitle timings.

2 Literature Review

2.1 Multilingual Subtitling Systems

MUltilingual Subtitling of multimediA content (MUSA)

MUSA (Piperidis, Demiros, and Prokopidis, 2004) combines speech recognition, advanced text analysis, and machine translation to help generate multilingual subtitles in English, French and Greek. Subtitling processes include tokenisation and subtitle text splitting, calculation of cue-in/cue-out timecodes, handcrafted rules, and shallow parsing. Subtitles must to some extent summarise the speech, as people can listen to more than they can read in the same time. MUSA's sentence compression module (Daelemans & Höthker, 2004) uses shallow-parsing information and handcrafted deletion rules to:-

- remove disfluencies such as repetitions introduced by hesitation
- replace part of the input sentence by a shorter paraphrase

eTITLE

eTITLE (Melero, Oliver and Badia, 2006) is a web-based multilingual subtitling service for the English, Spanish, Czech and Catalan. The system is configured as a distributed environment, so that different translation memories and machine translation modules can be located on different computers. A script and video file is inserted before any processing takes place. Like MUSA, eTITLE's translation module integrates translation memories with machine translation, because their shortcomings and advantages are complementary:

> *Since the output [of translation memories] is originally generated by humans, the typical errors and noisy output that MT systems sometimes produce are avoided. However, they are less robust than MT. (p2., Melero, Oliver & Badia, 2006).*

2.2 Speech Recognition

2.2.1 HMM-based Speech Recognition

Myers and Whitson's (1995) implementation of Rabiner's statistical speech recognizer (1989) is described as a set of programs that:

1. converts audio/wave files to sequences of multi-dimensional feature vectors. eg. DFT (discrete Fourier transform), PLP (Perceptual Linear Prediction), etc.
2. quantizes feature vectors into sequences of symbols eg. VQ (Vector Quantization)
3. trains a model for each recognition object (ie. word, phoneme) from the sequences of symbols. e.g. HMM)
4. constrains models using grammar information.

Phonemes are the standard unit of sound used, being the smallest unit of speech that can differentiate one word from another. English phonemes include 'ue' sound in 'moon', 'blue', 'grew' and 'tune', and 'f' in 'field' and 'photo'.

2.2.2 Grapheme-based Speech Recognition

A grapheme-based system (Killer, Stüker, and Schultz, 2003) can be used for languages with strong grapheme-phoneme relations. The advantage for a less-resourced language with lack of trained phoneticians (Williams, 2008) is that audio files and transcripts are sufficient data to train the system. Tests reveal the relative strength of phoneme-grapheme relations in different languages:

	Word Error Rate (WER)		
Language	**English**	**German**	**Spanish**
Phoneme	12.7%	17.7%	24.5%
Grapheme	19.1%	17.0%	26.8%

Table 1: Phoneme-based vs Grapheme-based recognition (p.3142, Killer, Stüker, and Schultz, 2003)

Canolfan Bedwyr's research on developing a text-to-speech synthesiser confirmed that "[i]n contrast to English, Welsh orthography is a reliable guide to pronunciation." (p.2, Williams, 1995). In addition, "Welsh, unlike English, does not exhibit stress-related vowel reduction or vowel lengthening." (p.2, Williams, 1995).

2.2.3 Speech Recognition and Disfluencies

Other research of relevance is the varying performance of speech recognition depending on the context of spoken text. Duchateau, Laureys, and Wambacq (p.1, 2004) found that:

> *the recognition accuracy of freely spoken language is quite poor when compared to that of dictated speech: while the word error rate(WER) for large vocabulary speaker-independent dictation is about 5%, the WER for spontaneous speech recognition ranges from 15% for broadcast news.. to 40% for meeting and telephone conversation transcription.*

2.2.4 Welsh Speech Recognition

There is no Welsh speech recognizer currently available. However, a significant output of Canolfan Bedwyr's (Welsh and Irish Speech Processing Resources) WISPR Project (Williams, Jones and Uemlianin, 2006) was a large grapheme-to-phoneme dictionary, suitable for use in a speech recognizer. The KTP adviser for this project is a speech recognition expert who was involved in Canolfan Bedwyr's WISPR project. We aim to develop a Welsh HMM-based speech recognizer that can be integrated into the automated subtitling system.

2.3 Translation

2.3.1 Rule Based Machine Translation (RBMT)

Apertium

Apertium is a sophisticated rule-based machine translation system that uses a sequence of finite-state automata:-

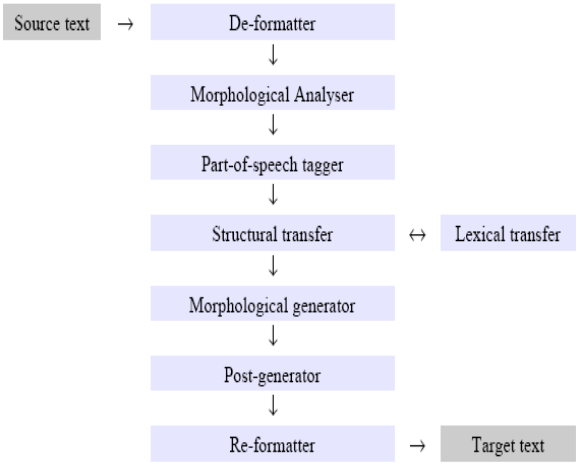

Figure 1: The eight modules of the Apertium system (p.4, Corbí-Bellot et al., 2005)

The deformatter extracts text from formatted documents and outputs a lemma, lexical category and morphological inflectional information. The HMM part-of-speech tagger is trained on representative source language text. The lexical transfer module outputs dictionary entries for target language words and multi-word units. The structural transfer module identifies chunks or phrases requiring additional grammatical processing to account for structural differences. The morphological generator inflects target language surface forms. The post-generator performs contractions and inserts apostrophes. The re-formatter provides formatting as per the original file.

2.3.2 Statistical Machine Translation (SMT)

Moses

Moses (Koehn et al., 2007) is one of the leading statistical-machine translation software available. The package includes corpus preparation (tokenisisation, converting to lowercase), statistical data (n-grams, statistically-derived word classes), machine translation steps (word and phrase alignment, translation), tuning and evaluation. A key attraction is that that the system is language independent. In practice, the success of SMT depends on the language pair and domain.

2.3.3 Translation Memories and Example Based Machine Translation (EBMT)

Armstrong et al. EBMT

Armstrong et al. (2007) researched using translation memories and Example Based Machine Translation (EBMT) for subtitling films. Notwithstanding user feedback that post-edited EBMT generated German subtitles of film dialogue displayed good colloquial phrases, had the correct tone and register and "felt like German", users preferred the output of rule-based Babelfish MT output over raw EBMT. The translation difficulties encountered may stem largely from developing a system for a broad subject-matter and spoken dialogue.

Leplus, Langlais and Lapalme

Leplus, Langlais and Lapalme (2004) developed a limited-domain translation memory and EBMT system for weather reports. They created a bilingual corpus from Environment Canada forecast reports in French and English. With the full memory (about 300,000 source sentences), 87% of sentences were found verbatim in the memory, and 89.5% always exhibited the same translation, possibly because these translations were generated by the rule-based METEO machine translation system.

Déjà Vu

Testun has some understanding of Atril's Déjà Vu, as it is the translation memory used by their Translation Department. Déjà Vu uses three types of resource:-

- translation memory of source and translated sentences,
- terminology database, and
- lexicon of translated word or multi-word terms from common words in files to be translated

With this information, Déjà Vu can:-

- search for source sentences that have been translated before
- search for similar sentences and highlight differences
- assemble a translation

This last function appears uses EBMT. If the source text contains the sentence:

Prometheus, the heavy equipment and engine manufacturer

the following French sentence from the translation memory would partly correspond:

Prometheus, the heavy equipment and engine producer

Prometheus, le producteur de matériel lourd et de moteurs

If the terminology database contains "producer" and "manufacturer", Déjà Vu can assemble a translation by replacing "*producteur*" with "*fabricant*".

The key to successful implementation would therefore be supplying the system with an extensive terminology database and lexicón

2.3.4 Welsh-English Machine Translation

English-Welsh machine translation systems have been developed – for example Phillips (2001) implemented a statistical system, Jones and Eisele (2006) tested the Pharaoh (Koehn, 2004) system (precursor to Moses), and Tyers and Donnelly (2009) adapted the rule-based Apertium system for Welsh-English translation. Both Apertium (Tyers and Donnelly, 2009) and Moses (Koehn et al., 2007) can produce excellent gist translations:-

Welsh original sentence	Mae Heddlu'r De yn ymchwilio i farwolaeth dyn 41 oed o Abertawe.
Apertium translation	South Wales Police is investigating death man 41 years old Swansea.
Moses translation	the south wales police investigation into the death of a man 41 years of age of abertawe.

Table 2: Apertium and Moses output for Welsh-English translation (Donnelly, 2009)

The output does not satisfy S4C's quality standards. However, Welsh-English machine translation is expected to improve – there is a project under the fifth Google Summer of Code to combine Moses and Apertium.

We would be most interested in using machine translation where output meets S4C's quality criteria and is acceptable to end-users.

2.4 Useful Techniques from Information Retrieval (IR)

IR techniques could play an important part in this project, both in the preparation of suitable data, and selection of appropriate sentences to be used as subtitles.

2.4.1 Term Extraction

2.4.1.1 Word Clustering

Word clustering techniques can be useful to form equivalence classes that can be used as interchangeable variables during language generation (Brown, 1999). For example, the sentence

> *John Miller flew to Frankfurt on December 3rd*

can yield the following pattern:

> *<firstname> <lastname> flew to <city> on <month> <ordinal>*

Words that have a similar context of surrounding words are deemed to belong to the same class. Brown's (1999) system looks at previous and succeeding words up to three positions apart. Equivalence is measured by cosine similarity. The standard formula for cosine similarity is:-

$$sim(s,y)=\cos(\vec{x},\vec{y})=\frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}=\frac{1\cdot1+1\cdot0+0\cdot1}{\sqrt{(1^2+1^2+0^2)*(1^2+0^2+1^2)}}$$

As a simplified example, if word **x** has as its context the preceding word **a** and following word **b**, and word **y** has context words **a** and **c**, a table can be formed as follows:-

	a	**b**	**c**
x	1	1	0
y	1	0	1

Table 3: Equivalence by cosine similarity

Brown adapted the cosine similarity formula to add increasing weight to increasingly adjacent words. An aligned word pair is made on the basis of dictionary entries, and must meet a certain threshold of similarity to join an existing bilingual cluster.

2.4.1.2 Multi-Word Units

Statistical measures can also be useful for extracting multi-word units. Based on the observation that words within multi-word units have high n-gram probability and low entropy, while there is considerable variance in words surrounding the multi-word units, Shimohata et al. (1997) devised an entropy-based formula for extracting multi-word terms:

$$H(str)=\sum_{i=1}^{n}-p(w_i)\log p(w_i)$$

To avoid uninteresting n-grams such as 'label for the', Merkel and Andersson (2000) compiled four stop-word lists:

- non-starters cannot begin a multi-word unit
- non-enders cannot end a multi-word unit
- prohibited words cannot be part of a multi-word unit
- ignored words are statistically insignificant and are ignored for entropy calculations

This 'linguistic-lite' approach is flexible and portable, which benefits less-resourced languages. The methodology may be adapted to use part-of-speech categories rather than word lists e.g. non-enders could be prepositions, articles, conjunctions and verbs.

2.4.2 Measuring Sentence Similarity

Finding a Welsh sentence in a parallel corpus that is similar to an input spoken sentence is an IR task. Jin and Barrière (2005) evaluated four formulae that measure sentence similarity by treating sentences as pure strings, stating that:

> *"The advantage of such an approach compared to more linguistically motivated approaches is that the system can quickly retrieve similar sentences from a large size corpus (over one million sentences), work well with ill-structured sentences, and work across different human languages." (p.1, Jin and Barrière, 2005).*

For an automated subtitling system, sentences need to be retrieved quickly to generate live subtitles. Also, syntactic analysis may not work with input sentences that, even when scripted,

are formulated for spoken rather than written well-formedness. The measures are also language-neutral which render them suitable for use with a less-resourced language.

In the formulae below, Q represents a query or input sentence, and S a sentence from the corpus. Three of the formulae evaluated are IR similarity-ranking functions:

- the Dice Coefficient (Hersh, 2003)

$$Dice(Q,S) = \frac{(2 * N_{common})}{(N_Q + N_S)}$$

- Vector Space Model (Cosine) (Salton and McGill, 1983), and

$$COSINE(Q,S) = \frac{\sum_{k=1}^{t}(w_{qk} \cdot w_{sk})}{\sqrt{\sum_{k=1}^{t}(w_{qk})^2 \cdot \sum_{k=1}^{t}(w_{sk})^2}}$$

- Lin's Information Theory Similarity (Aslam and Frost, 2003)

$$IT - Sim(Q,S) = \frac{2\sum_{w \in Q \cap S} \log \pi(w)}{\sum_{w \in Q} \log \pi(w) + \sum_{w \in S} \log \pi(w)}$$

The fourth formula evaluated was BLEU (Papineni et al., 2002), used to compare the output of machine translation with reference translations:

$$\text{Log BLEU} = \min(1 - r/c, 0) + \sum_{n=1}^{N} w_n \log p_n$$

Each formula was tested on English, French and Chinese text, and four sentences were retrieved for each input sentence. The output of the retrieved sentences were evaluated by human testers:

Grade	Explanation
4	The sentence exactly matches the input
3	The sentence provides enough information about the whole input
2	The sentence provides information about some part of the input
1	The sentence provides no information about the input

Table 4: Accuracy grades (p.6., Jin and Barrière, 2005)

The scores across languages were:

Algorithm	Average similarity score received (Highest score is 4)
Dice	2.73
Cosine	2.75
Lin	2.73
BLEU	2.64

Table 5: Average similarity score for different algorithms across languages (p.6, Jin and Barrière, 2005)

Ranking of the four sentences retrieved by each formula were compared with human ranking of the same sentences:

Algorithm	Percentage of agreement with human rating
Dice	100%
Cosine	93%
Lin	67%
BLEU	80%

Table 6: Average similarity score for different algorithms across languages (p.6, Jin and Barrière, 2005).

3 Project Plan

A suitable starting point for an automated Welsh-English subtitling system would be weather bulletins. S4C currently broadcasts weather bulletins six times a day during weekdays with at least three bulletins on Saturday and four bulletins on Sunday. The bulletins are not currently subtitled, but they will be by the end of the year to meet S4C's subtitling targets. The contract for subtitling weather bulletins will specify S4C's quality requirements, but not the implementation. For Testun, employing subtitling staff for short bulletins at otherwise non-busy times would be expensive. An automated subtitling system could help the company offer a competitive subtitling service for this programme.

Here is a summary of the project plan:

- prepare script and audio files for the script-audio alignment module
- build a suitable bilingual parallel corpus
- extract or compile domain-specific terminologies
- integrate translation memories and summarizer

- integrate a Welsh speech recognizer
- testing, debugging, refinement and rollout

The client and developer's view of the system are represented in figures 2 and 3:

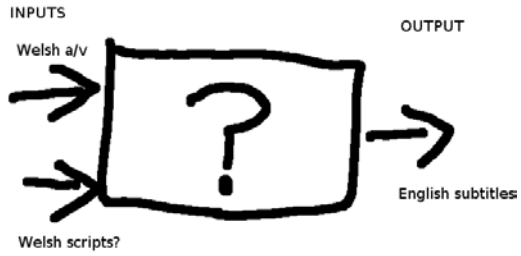

Figure 2: Client view of Testun's automated subtitling system

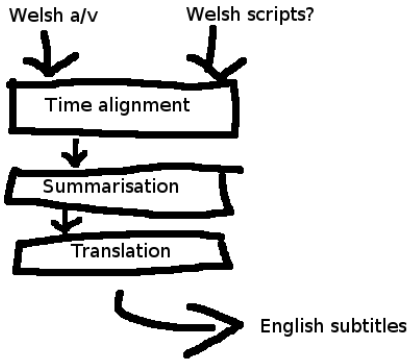

Figure 3: Developer view of Testun's automated subtitling system

3.1 Preparation of Script and Audio Files

Testun uses Sysmedia's WinCAPS subtitling system for its subtitling work for S4C. Sysmedia has also developed the SpeechFollower script-audio module. This module can be used for any language. Audio files to train the system should contain a range of text spoken by six different speakers, including men and women, and people with various regional accents. This speaker-independent system can take as input the speech of actual presenters rather than re-speakers. SpeechFollower handles timecodes to ensure that subtitles are generated in good time and at a readable pace.

3.2 Build a Bilingual Corpus

Testun not only provides subtitling, but also Sbectel teletext services for S4C – textual information pages available for view on a TV. Teletext weather bulletins are available twice a day in English and Welsh, and are exact translations. Testun has three months of archive teletext bulletins available for a parallel corpus. Teletext pages are concise, and the short sentence structures are highly suitable for use as subtitles.

3.3 Extraction of Domain-Specific Terminologies

Preliminary analysis of bulletins indicate that weather, temporal and location terms would be useful. We will experiment with the clustering and term extraction techniques outlined in section 2.4.1 to extract interchangeable words and multi-word units such as 'sunny intervals' and 'over the course of the week'. Following alignment and term categorisation, new sentences can be generated to expand the parallel corpus.

3.4 Integration of Translation and Summarizer Components

A comprehensive literature review and testing of machine translation and translation memory systems will be conducted. The component must be capable of seamless integration with WinCAPS subtitling software. We will also look into the feasibility of matching scripted sentences with Welsh sentences from the teletext corpus, using a sentence similarity metric (see section 2.4.2). Text summarisation may involve selecting sentences for which the closest match is found and/or which score highly for domain-specific terms. The equivalent English sentences can be output as a subtitle with confidence in their suitability and grammaticality. However, the results will need to be evaluated carefully to ensure coherent sequence and no significant loss in information. Alternatively, an English text summarizer may be used.

3.5 Integration of a Welsh Speech Recognizer

Some presenters do not prepare scripts and others are inclined to ad-lib than follow their

scripts faithfully, rendering the script-speech alignment module inoperable. It will therefore be necessary to integrate a Welsh speech recognizer. We may use a freely available speech recognition toolkit to develop our own time-alignment software, such as the open-source speech recognition toolkit Sphinx and/or the free-of-charge HTK (Hidden Markov Model Toolkit). Practical experience of working with the SpeechFollower module will indicate the procedures necessary to achieve time-alignment.

3.6 Testing, debugging, refinement and rollout

While Testun's Company Director and the Head of Subtitling Unit will be involved in all stages of requirements, design and evaluation of the system, a thorough evaluation of overall performance will be undertaken at this stage. Criteria for evaluation will include use of resources and speed. Extensive evaluation of overall performance will also be undertaken. Standard measures such as WER (Tillmann et al. 1997) and BLEU (Papineni et al., 2002) will be used to measure speech recognition performance and translation performance respectively. Human evaluation will be provided by Testun's subtitling staff.

4 Expected Results

We hope to develop usable resources as soon as possible. Some examples are:

- Shortcuts for subtitlers. Shortcuts are abbreviated terms that a subtitler can use within WinCAPS to aid fast typing. We will extract Welsh phrases used frequently in the forecasts and create English shortcut dictionaries for WinCAPS.
- Audio and data files for training the speech-audio alignment module
- A bilingual parallel corpus of weather forecasts
- Single and multi-word domain-specific, interchangeable, weather, temporal and location terms.
- A sentence similarity matcher and/or summarizer
- A Welsh speech recognizer

5 Conclusions

This paper has outlined the scope and challenges of a two-year KTP project for the development of automated English-Welsh subtitling software. Prior research indicate ways of mitigating the challenges, so that limited but practical systems can be developed. There is an incentive to get a good working system in use in the very near future, while providing opportunities for incremental development. The project will be deemed successful if careful and appropriate use is made of the automated subtitling system, helping the Welsh language TV industry increase services for social inclusion and accessibility.

Bibliography

Apertium <http://www.apertium.org/> Viewed 15 July 2007

Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D., O'Hagan, M., 2007. Leading by Example: Automatic Translation of Subtitles via EBMT. In *Perspectives, 14:3,163 — 184*

Aslam, J. Frost, M. 2003 An Information theoretic Measure for Document Similarity. In *Proceedings of the 26the Annual International ACM SINGIR Conference on Research and Development in Information Retrieval (ACM Press) 449-450*

Atril Déjà Vu X <http://www.atril.com/> Viewed 15 July 2007

Brown, R. D. 1999. Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22-32.

Canals R., Esteve, A., Garrido, A., Guardiola, M. I., Iturraspe-Bellver, A., Montserrat, S., Pérez-Antón, P., Ortiz, S., Pastor, H., and Forcada, M. 2001. InterNOSTRUM: a Spanish-Catalan Machine Translation System. In *Machine Translation Review, 11:21-25.*

Cancelo, P., 2000, [Reseña a] Herramientas Mágicas / Word Magic Tools de Word Magic Software (Word Magic Tools 2000

Deluxe 2.1.). In *Revista de Lexicografía, VI (1999-2000)*, pages 235-238

Corbí-Bellot, A.M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor A., Sarasola, K. 2005. An open-source shallow-transfer machine translation engine for the romance languages of Spain. In *Proceedings of the European Associtation for Machine Translation, 10th Annual Conference* (Budapest, Hungary, 30-31.05.2005), pages 79—86

Daelemans, W. & Höthker. A .2004. Automatic Sentence Compression in the MUSA project. In *Languages & The Media*, Berlin

Duchateau, J., Laureys, T., Wambacq, P. 2004. Adding Robustness to Language Models for Spontaneous Speech Recognition. In *the Proceedings of Robust2004, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*

Eurfa <www.eurfa.org.ok> Viewed 15 July 2009

Hersh, W. 2003. Information Retrieval: A Health & Biomedical Perspective (Second Edition, Springer-Verlag), chap. 8.

Hidden Markov Model Toolkit (HTK) <http://htk.eng.cam.ac.uk/> Viewed 15 July 2009

Huggins-Daines, D. The CMU Sphinx Group Open Source Speech Recognition Engines. <http://cmusphinx.sourceforge.net/html/cmusphinx.php> Viewed 15 July 2009

Jin, Z. Barrière, C. 2005. Exploring Sentence Variations with Bilingual Corpora. In *Corpus Linguistics 2005 Conference*, Birmingham, United Kingdom

Jones, D. and Eisele, A. 2006. Phrase-based statistical machine translation between English and Welsh. Strategies for developing machine translation for minority languages in *5th SALTMIL workshop on Minority Languages, LREC-2006,* Genoa

Killer, M., Stüker, S., and Schultz, T. 2003. Grapheme Based Speech Recognition. In *Proceedings of the Eurospeech.* Geneva, Switzerland.

Koehn, P. 2004. Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas, Lecture Notes in Computer Science. AMTA,* Springer.

Koehn, P., Hoang, H., Birch, A., Callison Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, demonstration session.*

Leplus, T., Langlais, P. and Lapalme, G., 2004. Weather Report Translation using a Translation Memory. In *AMTA 2004: 154-163*

Merkel, M., Andersson, M. 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings of 2000 conference user-oriented content-based text and image handling* (pages 737-746), Paris, France.

Melero, M. Oliver. A and Badia, T. 2006. Automatic Multilingual Subtitling in the eTITLE Project. in *Proceedings of ASLIB Translating and the Computer 28*

Microton Intelligent Software: <http://www.eurotran.cz/> Viewed 15 July 2007

Myers, R. And Whitson, J. 1995. Hidden Markov Model for automatic speech recognition: <http://read.pudn.com/downloads71/sourcecode/graph/254695/hmm-1.03/README.hmm__.htm> Viewed 15 July 2007

Papineni, K. Roukos, S. Ward, T. Zhu, W. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 311-318*

Piperidis, S., Demiros, I.,Prokopidis. P. 2004. Multimodal multilingual information processing for automatic subtitle generation: Resources, Methods and System Architecture (MUSA). In *Languages & The Media,* Berlin

Piperidis, S. Demiros, I. Prokopidis, P. 2006. Infrastructure for a multilingual subtitle generation system. in *Linguistics in the*

Twenty First Century, pages 369-378. Cambridge Scholars Press

Planas, E. SIMILIS. Second-Generation Translation Memory Software. in *ASLIB CONFERENCE, Translating and the Computer 27 Conference Programme*. <http://www.aslib.co.uk/conferences/ programme27.html> Viewed 15 July 2007

Phillips, J. D. 2001. The bible as a basis for machine translation. In *Proceedings of PACLing 2001*.

Prys, D., Williams, B., Hicks, B., Jones, D., Chasaide, A.N, Gobl, C., Berndsen, J., Cummins, F., Chiosáin, M. N., McKenna, J., Scaife, R., Dhonnchadha, E. U. , 2004. WISPR: Speech Processing Resources for Welsh and Irish. In *Pre-Conference Workshop on First Steps for Language Documentation of Minority Languages, 4th Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal.

Rabiner, L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*.

Salton, G. McGill, M. 1983. Introduction to Modern Information Retrieval, McGraw-Hill Book Company.

Systran < http://www.systran.co.uk/> Viewed 15 July 2009

Sysmedia <http://www.sysmedia.com/> Viewed 15 July 2009

Tr-AID Translation Memory <http://www.ilsp.gr/traid1_eng.html> Viewed 15 July 2009

Tillmann, C., S. Vogel, H. Ney, A. Zubiaga & H. Sawaf: 1997, Accelerated DP based search for statistical translation. In *Proceedings of 5th EUROSPEECH*, pp. 2667–2670.

Tyers, F. M. and Donnelly, K. 2009. *apertium-cy - a collaboratively-developed free RBMT system for Welsh to English*. In The Prague Bulletin of Mathematical Linguistics No. 91, pp. 57–66

Donnelly, K. 2008. Cyfieithu awtomatig a'r Google Summer of Code. <http://ilazki.thinkgeek.co.uk/~donnek/extras/pr2009_05_12.php> Viewed 15 July 2009

Williams, B., 1995. Text-to-speech synthesis for Welsh and Welsh English. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH 95)*, Madrid, Spain.

Williams, B., Jones, R. J. and Uemlianin, I. 2006. Tools and resources for speech synthesis arising from a Welsh TTS project. In *Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, 24-26.

Williams, B and Jones, R. 2008. Acquiring Pronunciation Data for a Placenames Lexicon in a Less-Resourced Language. In *Proceedings of the 6th LREC (International Conference on Language Resources and Evaluation).*