[*Translating and the Computer 28*, November 2006 (London: Aslib, 2006)]

Improving the Quality of Automated DVD Subtitles via Example-Based Machine Translation

Stephen Armstrong Andy Way
National Centre for Language Technology
School of Computing
Dublin City University
Dublin 9, Ireland
{sarmstrong,away}@computing.dcu.ie}

Colm Caffrey Marian Flanagan Dorothy Kenny Minako O'Hagan
Centre for Translation and Textual Studies
SALIS
Dublin City University
Dublin 9, Ireland
{colm.caffrey,marian.flanagan,dorothy.kenny,minako.ohagan}@dcu.ie

October 16, 2006

Abstract

Denoual (2005) discovered that, contrary to popular belief, an EBMT system trained on heterogeneous data produced significantly better results than a system trained on homogeneous data. Using similar evaluation metrics and a few additional ones, in this paper we show that this does not hold true for the automated translation of subtitles. In fact, our system (when trained on homogeneous data) shows a relative increase of 74% BLEU in the language direction German-English and 86% BLEU English-German. Furthermore, we show that increasing the amount of heterogeneous data results in 'bad examples' being put forward as translation candidates, thus lowering the translation quality.

1 Introduction

The demand on subtitlers to produce high-quality subtitles in an ever-diminishing space of time is at a record high, with many believing that a technology-based translation approach is the way forward (O'Hagan. 2003; Carroll, 1990; Gambier. 2005). Following on from recent research (Armstrong et al., 2006a,b) where we documented our motivations for using Example-Based Machine Translation (EBMT) in the Subtitling domain and produced some rudimentary translations, we have now come to the stage of improving the output quality of our system. EBMT relies heavily on a parallel aligned corpus, on which the system is trained. The question then arises: what type of

corpus will improve translation quality the most? A language-specific corpus, or a corpus containing out-of-domain data?

This paper aims to investigate whether a correlation exists between the quality of DVD subtitles and the corpus used to train the system. We present a modular Machine Translation system, newly developed at the NCLT in Dublin City University (Stroppa et al., 2006), which we use to translate subtitles from English into German by way of EBMT. The system was loaded with separate sets of both homogeneous data (ripped subtitles) and heterogeneous data (parliamentary proceedings), and a number of experiments were conducted to determine which dataset produced the highest quality output.

The remainder of this paper is structured as follows: In Section 2, we briefly discuss recent research in the area of homogeneous and heterogeneous data relevant to EBMT. We give an overview of EBMT and the Marker Hypothesis in Section 3. Section 4, introduces the system and details the chunking, chunk alignment, and translation processes. In Section 5, we present the different types of evaluation conducted and discuss the results the system achieved when loaded with the different training datasets. Finally, we conclude the paper with a summary of the results from our evaluation and give an outlook on possible future research in this area.

2 Heterogeneous versus Homogeneous Data

With almost all research in MT today being carried out using corpus-based techniques, it is strange to note that there has been little study into the effect the training-corpus has on the final output of the system. Up until recently it was assumed that corpus-based MT systems achieve better results when trained with homogenous data. Denoual (2005) set out to reassess this general assumption, and discovered that, contrary to this belief, his system yielded better results when trained on heterogeneous data, compared with equal amounts of homogeneous data. Using the BTEC corpus (a multi-lingual speech corpus comprised of tourism related sentences) he randomly extracted 510 Japanese sentences and used these as input to the system. The system was then trained on increasing amounts of data from the remainder of the corpus, and automatic evaluation metrics (BLEU, NIST and mWER) were relied on to estimate the translation quality of the output produced by the system. Based on these three measures, he shows that for increasing amounts of data, translation quality improves across the board. More notably, when trained on the random heterogeneous data, translation quality is found to be either equal or higher than when using homogeneous data for training.

Denoual's findings prove true for larger amounts of data, but when trained on relatively small amounts (29,000 sentences and less), translation seemed to be of higher quality using the homogeneous data (based on NIST scores). No reason is given for this and it is unclear whether this cut-off point can be generalised for other types of corpora other than the sets he used during his experiments.

Obviously the nature of the data used to train a system will have implications on translation quality; however, one also has to take into account the nature of the data that will be used as input

to the system. Subtitles can appear very different from text in other domains; a quick glance at the statistics for our homogeneous corpus shows that sentences are much shorter compared with sentences from the Europarl corpus (see Section 5.1, Table 2). Even this one simple statistic suggests that we might be better off using a corpus of subtitles to train the system. As no previous research has been carried out with respect to the specific task of translating subtitles using a corpus-based approach, we believe that you cannot generalise that either homogeneous or heterogeneous will yield better results, thus warranting its own investigation.

3 EBMT and the Marker Hypothesis

The approach we take to the automatic translation of subtitles is Example-Based Machine Translation (EBMT). This is based on the intuition that humans make use of previously seen translation examples to translate unseen input. The system is trained on an aligned bilingual corpus, from which 'examples' are extracted and stored. During translation, the input sentence is segmented, and its constituents are matched against this example-database, with the corresponding target language examples being recombined to produce the final output.

Even though EBMT draws some parallels with Translation Memory there is one essential difference: TM software needs a human present at all times during the translation process, and does not translate automatically. EBMT, on the other hand, is an essentially automatic technique; having located a set of relevant examples, the system recombines them to derive a final translation, rather than handing them over to the human to decide what to do with them. Another major benefit of EBMT is that search goes beyond sentence-level, where subsentential examples are obtained, meaning we do not miss out on matches which may not be seen by looking at the sentence as a whole. Recently, the two paradigms are becoming more and more similar (Simard and Langlais. 2001), with second generation TM systems adopting a subsentential approach to extracting matches and postulating a translation proposal based on these matches.

3.1 Marker-Based Chunking

As mentioned in Section 3.2, the input along with the source-target training corpus has to be 'chunked' in order to obtain subsentential examples. The Marker Hypothesis (Green, 1979) states that 'all natural languages are marked for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes which appear in a limited set of grammatical contexts and which signal that context'. We have carried out several experiments (Way and Gough, 2005; Stroppa et al., 2006; Groves and Way, 2006) using this idea as the basis for the chunking component of our EBMT system, and found it to be a very efficient way of segmenting source and target sentences into smaller chunks. A set of closed-class (or marker) words, such as determiners, conjunctions, prepositions, and pronouns, are used to indicate where one chunk ends and the next one begins (Table 1), with the constraint that each chunk must contain at least one content (non-marker) word.

To make this process a little clearer, let us look at the following English-German example in (1):

Determiners	‹DET›
Quantifiers	‹Q›
Prepositions	‹P›
Conjunctions	‹C›
WH-Adverbs	‹WH›
Possessive Pronouns	‹POSS-PRON›
Personal Pronouns	‹ PERS-PRO›
Punctuation	‹PUNC›

Table 1: Some of the tags used during the chunking phase

(1) Darling, I'm sorry but I've lost my key
→Mein Guter, es tut mir leid aber ich habe meinen Schlüssel verloren

For the first step we automatically tag each closed-class word with its marker tag, as in (2):

(2) Darling ‹PUNC› , ‹PERS-PRO› I am sorry ‹CONJ› but ‹PERSJRO› I've lost ‹POSS-PRO› my key
→Mein Guter ‹PUNC› , ‹PERS_PRO› es tut ‹PERS.PRO› mir leid ‹CONJ› aber ‹PERS_PRO› ich habe ‹POSS.PRO› meinen Schlüssel verloren

As every chunk must contain at least one non-marker word, we just keep the first marker tag when multiple marker-words appear alongside each other and discard the rest (3):

(3) Darling ‹PUNC› , I am sorry ‹CONJ› but I've lost ‹POSS-PRO› my key
→Mein Guter ‹PUNC› , es tut ‹PERS-PRO› mir leid ‹CONJ› aber ich habe ‹POSS_PRO› meinen Schlussel verloren

3.2 EBMT - an Example

The task for the EBMT system is to translate the input sentence in (4) given the aligned data in (5) as its training corpus.

(4) Ich wohne in Paris mit meiner Frau

(5) Ich wohne in Dublin ↔ I live in Dublin
Es gibt viel zu tun in Paris ↔ There's lots to do in Paris
Ich gehe gern ins Kino mit meiner Frau ↔ I love going to the cinema with my wife

The data is then chunked (based on the Marker Hypothesis), with useful chunks and their target-language partners being extracted and stored for later use (6) and less useful chunks being cast

aside. These useful chunk pairs are identified using a range of similarity metrics (see Section 4.3.1).

(6) Ich wohne ↔ I live
in Dublin ↔ in Dublin
Es gibt viel ↔ There's lots
zu tun ↔ to do
in Paris ↔ in Paris
Ich gehe gern ↔ I love going
ins Kino ↔ to the cinema
mit meiner Frau ↔ with my wife

We start the translation process by searching the German side of the original corpus in (5) to see if it contains the whole sentence. It does not, so we chunk the input sentence into smaller constituents (7) using the same hypothesis for segmenting the original corpus, and search for these in the corpus of aligned chunks (6).

(7) Ich wohne
in Paris
mit meiner Frau

Having found these chunks in our database, they are recombined by the decoder (see Section 4.4) to produce the final translation in (8):

(8) I live in Paris with my wife

4 System Architecture

We use the MaTrEx (Machine Translation using Examples) system to produce the output used in our experiments in Section 5. The system is a corpus-based MT engine, and is designed in a modular fashion, allowing the user to extend and re-implement modules at ease. The main modules are as follows:

- Word Alignment Module: takes as input an aligned corpus, and produces a set of word alignments;
- Chunking Module: takes as input an aligned corpus, and produces a corpus of source and target chunks:
- Chunk Alignment Module: takes in source and target chunks, and aligns them sentence by sentence;
- Decoder: searches for a translation using the original aligned corpus and derived word and chunk alignments;

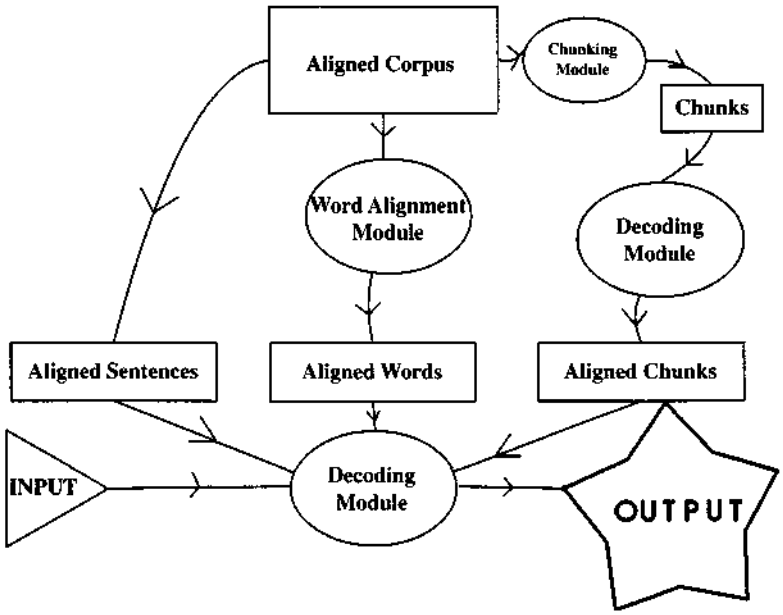

Figure 1: The system Architecture

4.1 Word Alignment

For word alignment we use the GIZA++ statistical word alignment toolkit, and following the refined method of Och and Ney (2003). extract a set of highly confident word-alignments from the original uni-directional alignment sets.

4.2 Chunking Module

Based on the Marker Hypothesis outlined in Section 3.1, we tag each source-target sentence in the training set with their corresponding marker tags. In total we use 452 marker words for English and 560 for German. Both sets of marker words are extracted from CELEX and edited manually to correspond with the training data. Examples of some of the tags used are shown above in Table 1.

4.3 Chunk Alignment

In order to determine alignments between chunks we use a dynamic 'edit-distance like' algorithm. Distances are calculated between each chunk in a sequence based on a combination of similarity metrics, and the most likely path is chosen between chunks. This algorithm is extended to allow for block movements, or jumps, following the idea introduced by (Leusch et al., 2006) and is incorporated to deal with potential differences between the order of constituents in English and German.

4.3.1 Computing Parameters for Chunk Alignment

Instead of using an Expectation-Maximization algorithm to estimate these parameters, as commonly done when performing word alignment (Brown et al., 1993; Och and Ney, 2003), we directly compute these parameters by relying on the information contained within chunks. In our experiments, we considered three main sources of knowledge: (i) word-to-word-to-word translation probabilities, (ii) distances based on chunk labels and (iii) distances based on the number of cognates per chunk. Word probabilities are taken from the process outlined in 4.1. Cognates are obtained by calculating the Lowest Common Subsequence (LCS), Minimum Edit-Distance and Dice Coefficient. As for Chunk labels, a simple matching process is used. All these sources of knowledge are combined using log linear model and are then used stored as a single parameter to determine the relationship between two chunks. This process is fully documented in Stroppa et al. (2006).

4.4 Decoding Module

Our example-based decoder is still under development (Groves, 2006) and not yet ready for use, so we decided to use the phrase-based decoder Pharaoh (Koehn, 2004) to search for and recombine target language candidates.

5 Experiments and Results

The aim of these experiments is to determine what yields better results for the translation of subtitles: training an EBMT system with specific data to a particular domain, or using data from a different source. We start by introducing the corpora used to train and test the system. We then go on to discuss the different experiments performed, which were twofold:

- Evaluation using automatic metrics
- Real-user Evaluation

5.1 The Corpora - Ripped Subtitles and the Europarl Corpus

For our study we first had to obtain sets of both aligned homogeneous and heterogeneous data for training and also gather together a corpus of sentences which would be suitable to test the system. To give a better idea of what we mean by homogeneous and heterogeneous data, we need to readdress the task that our EBMT system is faced with: the translation of subtitles. Subtitles themselves may come from a wide variety of scenes, across many different genres of movie and television. Although the type of dialogue used throughout a movie is mainly up to the discretion of the director and script writers, the subtitler has a less free role, and often has to conform to certain constraints, resulting in subtitles sharing certain similarities with a controlled language. [1]

[1] Simpler syntactic structures (canonical forms) are often preferred as they tend to make sentences shorter, thus more easily and quickly understood. Punctuation also differs greatly, and the subtitler must follow a number of rules which are not necessarily the same in natural language.

We make the assumption that a good example of homogeneous data would be to collect a set of subtitles along with their human translations. For our heterogeneous data we chose to extract at random sentences from the Europarl corpus, as we have used this data for several experiments at the NCLT and shown that the MaTrEx system performs well when trained and tested on such data. We found the best way to obtain our homogeneous data was to build up a collection of DVDs which included English and German subtitles, and then 'rip' these subtitles from the DVD into text format. DVD subtitles are stored as images and are blended into the video during playback, however it is possible to convert these images to text using the freely available software SubRip.[2]

Overall we ripped over 42,000 sentences of subtitles, aligned these based on techniques similar to those outlined in the chunk alignment process (see Section 4.3), and checked the resultant aligned sentences by hand to ensure accuracy. 40,000 of these sentence pairs were randomly selected and used as training data for the system, with the remaining 2,000 sentences being used as the test set. For our heterogeneous corpus we took a random sample of 40,000 sentences of English and their German equivalents from the Europarl corpus. Statistics for these data sets are shown in Table 2.

NUMBER OF SENTENCES, TYPES AND TOKENS

	sentences	tokens	types	ttr	sttr
Homogeneous Data	40000	226792	12399	5.47	39.62
Heterogeneous Data	40000	813297	19405	2.39	46.17
Test Data	2000	12197	2487	20.39	44.52

Table 2: Analysis showing the number of sentences, tokens and types, along with the type-token ratio (ttr) and standard type-token ratio (sttr) for the English training and test data

5.2 Automatic Evaluation

For this evaluation we used all 2000 sentences from the test set as input to the system and used their translation pairs as reference translations. We extracted sets of 10K, 20K, 30K and 40K sentences from both our homogeneous and heterogeneous corpora, used this data to train the system in separate experiments and estimated the impact these different datasets had on the output using a number of standard automatic evaluation metrics, namely:

- BLEU (Papineni et al., 2002) - Bounded between 0 and 1, where a higher score indicates a better translation. The geometric mean of the *n*-gram precisions is calculated with respect to a set of reference translations;
- NIST (Doddington, 2003) - Has a lower bound of 0, but no upper bound, where higher scores indicate a better translation. Variant of BLEU but is based instead on the arithmetic mean of weighted *n*-gram precisions in the output with respect to a set of reference translations;
- WER (Och, 2003) - Bounded between 0 and 1, where a lower score indicates a better translation. WER or Word-error rate is the edit distance in words between the system output and the reference translations.

[2]SubRip uses a similar technique to the optical character recognition (OCR) software used by scanners, where, with the help of the user, the images of the characters stored in the subtitle streams are converted into raw text.

5.2.1 Results for German → English

AUTOMATIC EVALUATION RESULTS

		BLEU	NIST	WER
10K	Homogeneous Data	0.1082	3.77	0.779
	Heterogeneous Data	0.0695	3.11	0.885
20K	Homogeneous Data	0.1166	3.96	0.776
	Heterogeneous Data	0.0740	3.21	0.876
30K	Homogeneous Data	0.1195	3.98	0.772
	Heterogeneous Data	0.0736	3.20	0.868
40K	Homogeneous Data	0.1287	4.08	0.761
	Heterogeneous Data	0.0737	3.21	0.865

Table 3: Automatic evaluation results for the test set when loaded with increasing amounts of heterogeneous and homogeneous data: German to English

The results for German-to-English translation are shown in Table 3. Note that as we increase the amount of homogeneous data, results show an improvement across the board, where the BLEU score increases by 20% when seeded with 4 times as much data. Increasing the amount of heterogeneous data does not seem to have as much of an effect on the translation quality, where we see a maximum increase in BLEU score of only 0.06%. Strangely translation quality is highest for 20K sentences of heterogeneous data, where 30K and 40K sets actually reduce the overall BLEU and NIST scores. What this suggests is that the system may be better off trained with less but more specific data, and that as we increase the amount of heterogeneous more 'bad examples' are introduced. Overall, the MT system scores 74% relative higher when loaded with homogeneous data.

5.2.2 Results for English → German

The results for the same experiment but in the opposite language direction are shown in Table 4. Although results are lower when compared with those in Table 3, we actually see a greater improvement when trained on our homogeneous data: the maximum BLEU score is 0.108 which when compared with the maximum for the heterogeneous, 0.058, suggests a relative increase of 86% BLEU. Again BLEU, NIST and WER scores improve with increments of homogeneous data, with 20K sentences again producing the best results for the heterogeneous data.

5.3 User Evaluation

Automatic evaluation is a quick, easy and often reliable way of getting an estimate of the translation quality of the output. However, none of the metrics mentioned above make use of linguistic information, and are mainly concerned with either counting @*n*-gram matches (BLEU and NIST). or calculating the edit distance between sentences based on words (WER). We propose a real-user evaluation, where we give a subject a set of machine-produced sentences and they are asked to give

AUTOMATIC EVALUATION RESULTS

		BLEU	NIST	WER
10K	Homogeneous Data	0.0769	3.22	0.912
	Heterogeneous Data	0.0517	2.53	0.991
20K	Homogeneous Data	0.0898	3.36	0.911
	Heterogeneous Data	0.0581	2.58	0.984
30K	Homogeneous Data	0.1040	3.55	0.891
	Heterogeneous Data	0.0529	2.59	0.989
40K	Homogeneous Data	0.1088	3.58	0.856
	Heterogeneous Data	0.0540	2.59	0.988

Table 4: Automatic evaluation results for the test set when loaded with increasing amounts of homogeneous and heterogeneous data: English to German

each sentence an intelligibility score (useful as a translation may not always resemble the source text, called translation by invention) and accuracy score (which relates to its closeness to some gold-standard translation). The intelligibility and accuracy scales are based on work by van Slype (1980) and by Nagao as described in Jordan et al. (1993) shown in Table 5.

	INTELLIGIBILITY SCALE
1	Easily Comprehensible
2	Comprehensible
3	Difficult to comprehend
4	Incomprehensible
	ACCURACY SCALE
1	Output sentence fully conveys the meaning of the source sentence
2	On the whole, the output sentence conveys the meaning of the source sentence
3	Output sentence does not adequately convey the meaning of the source sentence
4	Output sentence does not convey the meaning of the source sentence

Table 5: The scales and the range of scores possible for the User Evaluation

From our test set we extracted 200 sentences at random, and split these into four groups of 50 sentences. These were used as input to the system, which we trained on the full 40K sentences of homogeneous and heterogeneous data. Five native speakers of English (with fluent German) were asked to evaluate the German-English output, with five native speakers of German (with fluent English) being given the English-German output to evaluate.[3] The participants in the evaluation were first given the output produced for their mother tongue and asked to score each sentence for

[3]According to Kenny (personal communication) "it should be possible to evaluate intelligibility without any reference to the source text, so accuracy should not come into it; a text can be completely intelligible but bear little resemblance to the source text, accuracy, on the other hand, should be ascertained independently from intelligibility".

intelligibility. They were then given the source language set, used as input to the system, and asked to compare this with the output to give an accuracy score for the translation.

5.3.1 Results

Table 5.3 Average intelligibility and accuracy scores for the system trained on homogeneous and heterogeneous data.

USER EVALUATION RESULTS

		Intelligibility	Accuracy
	German - English		
40K	Homogeneous Data	2.45	2.98
	Heterogeneous Data	2.51	3.05
	English - German		
40K	Homogeneous Data	2.2	2.65
	Heterogeneous Data	2.7	2.8

Table 6: shows average Intelligibility and Accuracy scores when trained on equal amounts of homogeneous and heterogeneous data

Based on these results the system achieves best results when translating in the direction English-German and trained on homogeneous data: it achieves an average intelligibility score of 2.2 and accuracy score of 2.65. Translating in the same language direction, but using heterogeneous data to train the system, intelligibility and accuracy scores are 19% and 5% relative lower respectively. Similarly to the automatic evaluation, these results show that the system tends to perform best when trained on our homogeneous corpus of subtitles.

Due to the subjective nature of this evaluation approach, and the relatively small sample size of participants, results may appear somewhat skewed, depending on a person's interpretation of the terms intelligibility and accuracy. A larger number of participants is needed to get a more accurate measure of the translation quality. Another, perhaps even more useful way of using this type of evaluation, might be during during the development phase of a system. If one were to present these results in a qualitative rather than quantitative framework, we could use this as a good means of identifying where the system is going wrong and where improvements could be made.

6 Conclusions and Future Work

Using a number of evaluation techniques (both automatic and manual) we demonstrated that for the task of translating subtitles, the type of corpus used to train the system has a significant impact on translation quality. In addition to this, we showed that our EBMT system performs consistently better when seeded with a corpus of homogeneous data. We also noted that larger amounts of heterogeneous data seemed to produce 'bad examples', ultimately resulting in bad translations. Increasing the amount of homogenous data improved results across the board. However, we only used 40K sentences to train the system so it would be interesting to see if that trend continues, or if there is some tipping point where translation quality peaks.

Apart from increasing the amount of data we will use to train the system, other future research will focus on improving the accuracy of our chunk alignment strategies. This is where the qualitative user evaluation should prove useful. A HMM-based alignment strategy is also being worked on. Furthermore, we would like to implement an example-based decoder and make use of generalised templates, which should allow for more flexibility in the matching process, hopefully improving coverage and quality.

References

Armstrong, S., Caffrey, C., and Flanagan, M. (2006a). Leading by Example: Automatic Translation of Subtitles Using EBMT. In *Languages and the Media Conference [Forthcoming Presentation],* Berlin, Germany.

Armstrong, S., Caffrey, C., and Flanagan, M. (2006b). Translating DVD Subtitles Using Example-Based Machine Translation. In *proceedings from MuTra - Audiovisual Translation Scenarios Conference. [In Press],* Copenhagen, Denmark.

Brown, P., Pietra, S. D., Pietra, V. D., and Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. In *Computational Linguistics,* pages 263-311.

Carroll, M. (1990). Subtitling: Changing Standards for New Media.

Denoual, E. (2005). The Influence of Example-data Homogeneity on EBMT Quality. In *Proceedings of the Second Workshop on Example-Based Machine Translation,* pages 35-42, Phuket, Thailand.

Doddington, G. (2003). Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference (HLT-02),* pages 138-145, San Diego, CA.

Gambier, Y. (2005). Is Audiovisual Translation the Future of Translation Studies? In *Keynote Speech Delivered at the Between Text and Image, Screen-Translation Conference,* Forli, Italy.

Green, T. (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behaviour,* pages 481-486.

Groves, D. (2006). *Hybrid Data Driven Models of Machine Translation.* PhD thesis, Dublin City University, Dublin, Ireland.

Groves, D. and Way, A. (2006). Hybridity in MT: Experiments on the Europarl Corpus. In *Proceedings of the 11th Conference of the European Association for Machine Translation,* Oslo, Norway.

Jordan, P. W., Dorr, B. J., and Benoit, J. W. (1993). A First-Pass Approach for Evaluating Machine Translation Systems. *Machine Translation,* 8(1-2) :49-58.

Koehn, P. (2004). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA-04,* pages 115-124. Washington, District of Columbia.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT evaluation using block movements. In *In Proceedings of EACL-06,* page 241248, Trento, Italy.

Och. F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of 41st ACL,* pages 160-167, Sapporo, Japan.

Och. F. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics,* volume 29(1), pages 19-51.

O'Hagan, M. (2003). Can Language Technology Respond to the Subtitler's Dilemma? - A preliminary study. In *Proceedings of Translating and the Computer 25, ASLIB.* London, England.

Papineni. K., Roukas, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL,* pages 311-318, Philadelphia. PA.

Simard, M. and Langlais, P. (2001). Sub-sentential Exploitation of Translation Memories. In *Proceedings of MT Summit VII: Machine Translation in the Information Age,* pages 335-339. Santiago de Compostela, Spain.

Stroppa, N., Groves, D., Sarasola, K., and Way, A. (2006). Example-based Machine Translation of the Basque Language. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas,* pages 232-241, Boston, MA.

van Slype, G. (1980). Bewertung des Verfahrens SYSTRAN für die maschinelle Sprachübersetzung bei der K.E.G. *Lebende Sprachen: Zeitschrift für Fremde Sprachen in Wissenschaft und Praxis,* 25:6-9.

Way, A. and Gough, N. (2005). Comparing Example-Based and Statistical Machine Translation. In *Natural Language Engineering,* pages 295-309, Oslo, Norway.