# Example-Based Controlled Translation

*Nano Gough*

School of Computing
Dublin City University
Dublin 9, Ireland

ngough@computing.dcu.ie

*Andy Way*

School of Computing
Dublin City University
Dublin 9, Ireland

away@computing.dcu.ie

**Abstract.** The first research on integrating controlled language data in an Example-Based Machine Translation (EBMT) system was published in [Gough & Way, 2003]. We improve on their sub-sentential alignment algorithm to populate the system's databases with more than six times as many potentially useful fragments. Together with two simple novel improvements—correcting mistranslations in the lexicon, and allowing multiple translations in the lexicon—translation quality improves considerably when target language translations are constrained. We also develop the first EBMT system which attempts to filter the source language data using controlled language specifications. We provide detailed automatic and human evaluations of a number of experiments carried out to test the quality of the system. We observe that our system outperforms *Logomedia* in a number of tests. Finally, despite conflicting results from different automatic evaluation metrics, we observe a preference for controlling the source data rather than the target translations.

## 1. Introduction

Research in Machine Translation (MT) has explored many different methods over the years, including rule-based, statistical and example-based models as well as hybrid and multi-engine approaches. Certain MT systems have been developed for particular sublanguage domains. Furthermore, since 1996, there has been a growing interest in controlled languages and their application in MT as demonstrated by the series of CLAW workshops on controlled language applications. These have sparked the development of both monolingual and multilingual guidelines and applications using controlled language (CL) for many languages.

Natural language grammars can be restricted in such a way that ambiguity and complexity is lessened or eliminated completely. Controlled languages are subsets of natural languages whose grammars and dictionaries have been restricted for this purpose. As well as aiding human comprehension of texts, CLs can also be used for improving the computational processing of text and potential benefits have been claimed for the integration of controlled languages with translation tools.

Until quite recently, however, the area of Controlled Translation has been largely ignored. Only a limited number of rule-based MT (RBMT) systems have been used to translate controlled language documentation, including Caterpillar's CTE and CMU's KANT system [Mitamura & Nyberg, 1995], and General Motors CASL and LantMark [Means & Godden, 1996]. However, such systems can be very complex and expensive to develop for controlled translation, as it is difficult to fine-tune such general-purpose systems to derive specific, restricted applications.

It is widely recognised that the use of traditional RBMT systems can lead to the well known 'knowledge- acquisition bottleneck'. It is also acknowledged that the use of corpus-based MT technology can overcome this problem. It is difficult, therefore, to comprehend why more work has not been done in the development of Example-Based MT (EBMT) systems for controlled language applications, especially when one considers that the quality of EBMT systems depends heavily on the quality of the reference translations in the system database—the more these are controlled, the better

the expected quality of translation output by the system.

Recently [Gough & Way, 2003] presented the first attempt at controlled translation using EBMT. In this paper, they attempted to control the output translations by incorporating in the system's memories target language strings written according to *Sun Microsystems'* controlled language guidelines. In this paper we improve on their method of extracting sub-sentential alignments. In re-running their experiments with our new method, we succeed in populating the system's databases with considerably more sub-sentential fragments and demonstrate a considerable increase in translation quality.

As is acknowledged in [Gough & Way, 2003], it is more usual to propose the use of CL as a means of controlling the input texts rather than the output translations. In this paper, therefore, we use our improved methodology on their training and test data to control the processing of the source language. In assessing the results of [Gough & Way, 2003] and our improvements for French-English, we compare our novel results for English-French using manual and automatic evaluation metrics, and comment on the relative success of controlling source and target texts in controlled translation using EBMT. We also compare the results achieved with an array of automatic evaluation metrics. Finally, it has been claimed in the literature [Carl, 2003; Schäler *et al.,* 2003] that EBMT systems should fare better than RBMT systems when confronted with controlled data. To provide some experimental backup to these insights, we provide results for the good on-line system *Logomedia*, and compare these with the results obtained for our system.

The remainder of the paper is organised as follows: in section 2, we describe relevant previous research in the area of controlled translation. In section 3, we present our EBMT system and the methodology used to derive controlled translations. In section 4, we use both automatic and manual evaluation metrics to assess our system based on the results of different experiments. Finally, we conclude, summarise our contribution to the area of controlled translation in particular, and to EBMT in general.

## 2. Controlled Translation

Recent research [Carl, 2003; Schäler *et al.,* 2003] has addressed the theme of controlled translation and outlined some theoretical requirements for the development of MT systems for use with CLs. With respect to controlled translation in a transfer-based system, there are three stages of processing: it is necessary to exert control over the source language, the transfer routines as well as the generation component. With the absence of control at any one of these stages, one cannot necessarily expect to produce a high-quality controlled translation.

While the theoretical issues of controlled translation have been addressed to a certain extent, the lack of sententially aligned texts conforming to sets of controlled language specifications is a major obstacle in the development of applications for controlled translation. Although controlled language specifications do exist for English (e.g. CTE or CASL) and French (e.g. GIFAS Rationalised French [Barthe, 1998]), there is no controlled bitext in existence for any language. Moreover, the difficulty surrounding this task comes to light when we consider that there is no guarantee that enforcing different sets of controlled language specifications on both source and target documents would ensure the production of a necessary and sufficient translation.

However, some efforts have been made to automate this process. For example, [Hartley *et al.,* 2001; Power *et al.,* 2003] approach this task with respect to multilingual natural language generation. Users are prompted by the system to build up a text in one language in a technical domain. Although they need to be an expert in the specific domain, no foreign language knowledge is required. Instead, multiple expressions of the same underlying input in various languages is facilitated. While this task may be tedious, the strings will conform exactly to a strictly defined controlled language.

[Bernth, 2003] seeks to constrain the output so as to facilitate speech-to-speech translation. Bernth explores parse trees to identify undesirable constructions and rewrite them with suitable substituted target text. This method is, however, unavailable to us, as the corpus we use does not contain such detailed structural representations. The transfer-driven MT system of [Yamada *et al*., 2000] constrains transfer rules to control the generation of the correct forms of politeness in Japanese given English input.

More relevant to our approach is the previous work in the area of controlled translation using EBMT. [Gough & Way, 2003] use a corpus of *Sun* documentation written according to CL guidelines to

constrain the translations of 'unconstrained' input. They translate the controlled English text using the on-line system *Logomedia,* selected as it was deemed to be the better of the three on-line MT systems tested in [Way & Gough, 2003]. It is acknowledged in [Gough & Way, 2003] that while this may not be controlled translation *per se* according to the definitions of [Carl, 2003; Schäler *et al*., 2003], they justify this approach given the lack of availability of both controlled input and output.

In this paper, we extend the work of [Gough & Way, 2003] in two ways: firstly, we apply a number of improvements to their method of deriving sub-sentential resources which lead to enhancements in the quality of the French-English translations produced. Secondly, we train our system on the data used in their experiments on English-French, allowing us to make controlled *analysis* the focus of the research. For both experiments, we provide detailed automatic and human evaluations. In order to test the hypothesis that EBMT should be better suited to the task of controlled translation than rule-based methods, we also provide a comparison with *Logomedia.* This allows us to assess and compare the effects of both controlled analysis and generation on translation using our EBMT system.

## 3. Marker-Based EBMT

The 'Marker Hypothesis' [Green, 1979] is a universal psycholinguistic constraint, which states that languages are 'marked' for syntactic structure at surface level by a closed set of specific lexemes and morphemes. The Marker Hypothesis has been applied in previous EBMT systems including METLA [Juola, 1994], *Gaijin* [Veale & Way, 1997], and the *wEBMT* system [Gough *et al*., 2002; Way & Gough, 2003]. The previous work on controlled EBMT [Gough & Way, 2003] was also based on this *'linguistics-lite'* approach. The Marker Hypothesis is used to segment the aligned <source, target> strings at a sub-sentential level. Individual sets of marker words are established for English and French, and assigned to categories <DET>, <PREP> etc. These are then used to segment the aligned sentences, in order to generate a marker lexicon. As an example, consider the strings in (1) appearing in the *Sun* documentation:

(1) La première partie du livre décrit les composants du bureau ⇒
The first part of the book describes the components of the desktop

In a pre-processing stage, the aligned sentences are traversed word by word. A new sub-sententially aligned fragment begins where a marker word is encountered and ends at the occurrence of the next marker word, subject to each chunk containing at least one non-marker (or 'content') word. From the sentence pair in (1), the tagged strings in (2) are generated:

(2) <DET> La première partie <PREP> du livre décrit <DET> les composants <PREP> du bureau

<DET> The first part <PREP> of the book describes <DET> the components <PREP> of the desktop

In the *wEBMT* system, marker chunks in the source map sequentially to marker chunks in the target, subject to their marker categories matching. This seemingly naïve approach proves quite effective and in this way, smaller aligned segments can be extracted from the existing sentence-level lexicon without recourse to any detailed parsing techniques. Given the tagged strings in (2), the marker chunks in (3) are automatically generated.

(3) <DET> La première partie : <DET> The first part
<PREP> du livre décrit : <PREP> of the book describes
<DET> les composants : <DET> the components
<PREP> du bureau : <PREP> of the desktop

Given the marker chunks in (3), further lexical information can be extracted. We assume that where a chunk contains just one non-marker word in both source and target, these words are translations of each other. For example, from the third pairing in (3), we can extract the 'word-level' translations in (4):

(4) <DET> les : <DET> the
<LEX> composants : <LEX> components

That is, content-word translations can be derived automatically and are stored in our word-level lexicon using the <LEX> tag.

Finally, by generalising over the marker lexicon we produce a set of marker templates. This is achieved by replacing the marker word by its relevant tag. From the examples in (3), we can produce the generalised templates in (5):

(5) <DET> première partie : <DET> first part
<PREP>livre décrit : <PREP> book describes
<DET> composants : <DET> components
<PREP> bureau : <PREP> desktop

These templates increase the robustness of the system and make the matching process more flexible. Now any marker word can be inserted after the relevant tag if it appears with its translation in the lexicon. This causes a considerable amount of overgeneration, and many thousands of candidate translations may be suggested for any particular string. Nevertheless, each translation is output with its probability using the method of [Way & Gough, 2003], who showed that the 'best' translation always occurred in the top 1% of proposed translations, thereby facilitating pruning of the vast majority of translation candidates produced.

As an example, assume that we want to translate the string *ces composants,* but the only relevant entry in the marker lexicon is *les composants*, as in (3). In this case, the string might not be translated. However, by means of the generalised templates a translation can be produced. The input string <*ces composants*> is generalised to <*DET*> *composants*, which can be matched to the relevant template in (5). The insertion of the translation pair <*ces*, *these*> is allowed, given that this translation pair is found in the word-level lexicon with the marker tag <DET>, and a translation is derived.

### 3.1 An Improved Sub-sentential Alignment Method

Using the sub-sentential alignment algorithm of [Gough *et al*., 2002; Way & Gough, 2003; Gough & Way, 2003], marker chunks in the source map sequentially to marker chunks in the target, subject to their marker categories matching. Our revisions of their algorithm enable much more data to be retained. We check that chunks are marked with similar tags, as in the original method, but also take into account lexical similarity. A base-dictionary created via *Logomedia* is used to check for word-equivalences between chunks. Those chunks having one or more words in common are considered more likely alignments. Cognates are also considered to increase the likelihood of chunk alignment. Finally, the position of chunks in source and target sentences is also taken into account—the more distance between two chunks, the less likely they are to align. While the original algorithm of [Gough *et al*., 2002; Way & Gough, 2003; Gough & Way, 2003] could only account for 1:1 alignments, the new algorithm allows for multiple chunks in the source or target to merge, thereby making 2:1, 3:1 etc. alignments possible. For example, using the original alignment algorithm, the <source, target> pair in (6) would not be considered for chunk alignment:

(6) <QUANT> each layer has <DET> a layer number
<QUANT> chaque couche a <DET> un nombre <PREP> de la couche

Using the improved alignment method, however, the aligned chunks in (7) can be produced:

(7) <QUANT> each layer has : <QUANT> chaque couche a
<DET> a layer number : <DET> un nombre de la couche

That is, the last two chunks in the French sentence are merged to align with the final chunk in English. Any interim marker tags (such as <PREP>, here) are deleted in this process.

In [Gough & Way, 2003], 1079 sub-sententially aligned segments were produced in this way. Those chunks which could not be aligned via this method were translated by *Logomedia*, and if the translation produced was contained in the original translation, the chunks were also aligned. This produced an additional 2082 alignments (3161 in total). Using our new improved sub-sentential alignment algorithm, we populate the system's memories with more than six times as many aligned chunks on the same data using the same language pair (French-English), with no recourse to *Logomedia*. As a further comparison, while only 18% of the sentence pairs proposed candidates for chunk selection in [Gough & Way, 2003], over 85% of sentence pairs throw up sub-sentential candidate fragments using our improved method.

There are two other differences between the work presented in [Gough & Way, 2003] and this research. The first is that we carry out some limited updating of the word-level lexicon in this approach, and secondly, we permit multiple word-level translations to be used in the translation process. In [Gough & Way, 2003], the system's word-level lexicon derived from the process in (4) was fixed, and the number of options for each translation at the lexical level was restricted to one. As we demonstrate in the next section, both these minor amendments improve translation quality considerably.

## 4. Translation Experiments and Evaluation

In this section we report on a number of experiments carried out to test the system. We use the same testset as [Gough & Way, 2003] in order to directly compare our revised alignment method with theirs. 3885 sentences were extracted from a *Sun* Translation Memory dealing broadly with the same language area (computer documentation) as the CL data, but not written according to CL specifications. [Gough & Way, 2003] chose the French input strings on the basis that each word contained in these strings existed somewhere in the training corpus. For each unique word in the corpus, if a word did not exist in the word lexicon via the marker hypothesis alignment process (cf. (4) above), the word was translated on-line by *Logomedia* and added to the word-level lexicon.

We translated each of the 3885 sentences from French-English and English-French. In the following sections, we present both automated and human evaluations of the translations produced by the system for these language pairs. As a baseline comparison, we also provide results for *Logomedia*. We comment on the results obtained, and discuss the relative merits of the automatic metrics used.

### 4.1 French-English: Controlling the Target Language

#### 4.1.1 Automatic Evaluation

[Gough & Way, 2003] calculated IBM Bleu [Papineni *et al*., 2002] scores for the translations produced by their system using the NIST MT Evaluation Toolkit [1]. They also calculated Bleu scores for *Logomedia* on the same testset of 3885 sentences. They reported that when automatic metrics are utilised, *Logomedia* appears to considerably outperform their EBMT system: the average score for their system over the entire testset is 0.0836 compared to an average score of 0.1637 for *Logomedia*.

**Table 1.** Comparing our revised EBMT system (French-English) with *Logomedia* using the IBM Bleu Automatic Evaluation Metric on a 3885 Sentence Testset

| Bleu Score | Our System | Logomedia | Gough /Way 03 |
|---|---|---|---|
| Average | 0.1204 | 0.1637 | 0.0836 |
| Best Doc. | 0.1504 | 0.2244 | 0.1473 |
| Worst Doc. | 0.0667 | 0.0825 | 0.0462 |
| Best Sent. | 1.0000 | 1.0000 | 0.9131 |

Incorporating our novel refinements to the sub-sentential alignment algorithm of [Gough & Way, 2003], and eliminating the use of *Logomedia* to generate sub-sentential alignments, we calculated Bleu scores for the English translations produced on the same data as used in [Gough & Way, 2003]. These are shown in Table 1. The results show that using the revised sub-sentential alignment method, we obtain a translation score 44% higher than the method of [Gough & Way, 2003]. The best score for a sentence is 1.000, where previously it was 0.9131. Scores for best and worst document also increase with our new method. Note, however, that the raw, unamended EBMT system continues to lag behind *Logomedia* somewhat.

In an effort to increase the Bleu score for our system, we apply two simple, novel improvements to the system. Initially, we isolate the words that occur more than 10 times in our test corpus (10% of total words). We manually correct any mistranslations of these words (64 translations corrected), resulting in a Bleu score of 0.1267, a 5% improvement on the new baseline score of 0.1204. Given the success of this adjustment, we opted to correct all those words (57 additional translations corrected) occurring more than once within the corpus (30% of total words). Again, the Bleu score increased to 0.1449, an improvement of 20% on the new baseline figure, and 73% better than the average Bleu score reported in [Gough & Way, 2003]. Nevertheless, *Logomedia* still outperforms our system.

Finally, we reviewed the algorithm producing the final translation. Initially, in an effort to increase translation speed, the number of options for each word translation was limited to one. However, in many cases more than one translation was available for each word. We therefore adjusted the algorithm to allow for a maximum of five possible word translations to be used. Following this alteration, the Bleu score rose to 0.17, an improvement of 104% over the average reported in [Gough & Way, 2003], and a 41.67% improvement on the baseline Bleu score of 0.1204.

[1] http://www.nist.gov/speech/tests/mt/mt2001/index.htm

Perhaps more noteworthy is the fact that the Bleu score for our new, improved system is higher than the average Bleu score reported for *Logomedia* on the same data. While *Logomedia* is a good, general-purpose system, for the first time it can be seen that an EBMT system might be able to outperform an RBMT system. Of course, our system is trained on data similar to that contained in the test data, but we nonetheless are encouraged by this result, especially given that [Way & Gough, 2003] demonstrated that for uncontrolled data, *Logomedia* outperformed their *wEBMT* system.

In addition, we calculate Precision and Recall figures using the tools[2] reported in [Turian *et al.*, 2003] for both the new results and those presented in [Gough & Way, 2003], as well as Word-Error and Sentence-Error rates. These results are presented in Table 2. Like the Bleu score, using Precision and Recall shows an improvement using our new sub-sentential alignment algorithm, in that Precision improves by 45.5% and Recall by 0.9%. The benefits of the improvements to our system are also clearly seen in the WER and SER rates.

**Table 2.** Summary of results in comparing our revised EBMT (French-English) system with *Logomedia* and [Gough & Way, 2003] using Automatic Evaluation Metrics on a 3885 Sentence Testset

| Experiment | Precision | Recall | Bleu | WER | SER |
|---|---|---|---|---|---|
| Alignment 1 [Gough/Way 03] | 0.1815 | 0.3183 | 0.0836 | 96.7 | 98 |
| Alignment 2 | 0.2641 | 0.3211 | 0.1204 | 88.7 | 96 |
| Top 10% words corrected | 0.2722 | 0.3252 | 0.1267 | 86.1 | 95 |
| Top 30% words corrected | 0.2756 | 0.3302 | 0.1449 | 84.0 | 93 |
| Additional word Translations | 0.3005 | 0.3646 | 0.1703 | 80.1 | 88 |
| Logomedia | 0.2617 | 0.3601 | 0.1637 | 96 | 98.1 |

Precision, Recall and WER/SER figures also demonstrate that we now outperform *Logomedia.* Our best performance (improved sub-sentential alignment, correcting 30% of lexical translations, and allowing max. 5 translations per word) outperforms *Logomedia* by almost 4% Precision and 0.45% on Recall.

#### 4.1.2 Manual Evaluation

While these results with additional automatic evaluation metrics confirm those derived via Bleu, we decided to perform a manual evaluation to seek further confirmation that our novel amendments were contributing to translation quality. Accordingly, we carried out a manual evaluation on the same 200 sentences randomly extracted from the larger testset in [Gough & Way, 2003]. Each translation was measured according to the notions of intelligibility and accuracy (or fidelity). Intelligibility decreases if grammatical errors, mistranslations and untranslated words are encountered. However, a completely intelligible string may be output by an MT system, which is not a true reflection of the input. Therefore, accuracy is used to measure how faithfully the MT system represents the meaning of the source string on the target side. We use the same four levels of intelligibility as in [Gough & Way, 2003], from 'Score 3: very intelligible (accurate translation, no syntactic errors)' to 'Score 0: unintelligible'. Similarly, as in [Gough & Way, 2003], accuracy is measured on a 5-point scale: from 'Score 4: very accurate (good translation, represents source faithfully)' to 'Score 0: inaccurate'. Two native speakers of English with good French language competence carried out the task of evaluating these translations produced. The results showed that there was far less disparity between our system and *Logomedia* than was reflected by the automatic evaluation. Following the application of the revised alignment algorithm and the integration of some novel adjustments to the system and its lexical resources, the same metrics were applied to manually evaluate the translations produced. The results for intelligibility are presented in Table 3.

**Table 3.** Comparing our EBMT system (French-English) with *Logomedia* and [Gough & Way, 2003] in a Human Evaluation: Intelligibility

| System | Score 0 | 1 | 2 | 3 | Exact Match |
|---|---|---|---|---|---|
| Our System | 4 | 12 | 46 | 126 | 12 |
| Logomedia | 2 | 21 | 40 | 123 | 14 |
| *Gough/Way 03* | 10 | 30 | 35 | 118 | 7 |

With respect to intelligibility, we achieve 1.5% more score 3 translations than *Logomedia*. This is an improvement from [Gough & Way, 2003], where Logomedia outperformed their system by 2.5%. Regarding unintelligible translations, 5% of the output strings in [Gough & Way, 2003] were considered unintelligible, which falls to 2% using our system. In fact, contrary to the system of [Gough & Way, 2003], overall our system appears to outperform *Logomedia* on this evaluation criterion:

[2] http://nlp.cs.nyu.edu/GTM/

for scores 2, 3 and exact match (i.e. adequately or very intelligible translation with no syntactic errors), we obtain 184 (92%) such translations, while *Logomedia* obtains just 177 (88.5%). As many of the sentences in our system are translated with recourse to the word-level lexicon, the changes made to this resource, together with the revised alignment algorithm has presumably increased the intelligibility of the output translations. This addresses one of the issues outlined in [Gough & Way, 2003] where the potential benefits of additional word alignments produced from the example-base were noted.

**Table 4.** Comparing our EBMT system (French-English) with *Logomedia* and [Gough & Way, 2003] in a Human Evaluation: Accuracy

| System | Score 0 | 1 | 2 | 3 | 4 | Exact Match |
|---|---|---|---|---|---|---|
| Our System | 2 | 6 | 18 | 36 | 126 | 12 |
| Logomedia | 9 | 27 | 27 | 31 | 92 | 14 |
| Gough/Way 03 | 9 | 30 | 19 | 42 | 93 | 7 |

The results for accuracy are given in Table 4. Although *Logomedia* produces more exact matches than our system, we outperform [Gough & Way, 2003] on this measure. Where our system scores highly is in Score 4 (very accurate) translations: we outperform both *Logomedia* and [Gough & Way, 2003] by about 17%. Overall, we outperform *Logomedia* with regard to accuracy: 87% of the translations produced by our system obtain a score 3, 4 or exact match, while only 68.5% of translations produced by *Logomedia* fall into one of these categories (cf. [Gough & Way, 2003], who score 71% on translations scoring 3, 4 or better for accuracy).

The results in Table 4 reinforce the opinion of [Gough & Way, 2003] that Bleu is a harsh evaluation metric. 126 of the translations produced by our system were considered correct in a human evaluation. However, because they differ in some way from the oracle translation, they are penalised in the automatic evaluation.

#### 4.1.3 Summary

For French-English, the automatic evaluation metrics show that our system outperforms *Logomedia* on the 3885 testset of strings used in [Gough & Way, 2003]: our Bleu score is 0.66% higher than for *Logomedia*; we outperform *Logomedia* by almost 4% on Precision and by 0.45% on Recall. All automatic evaluation metrics (including WER and SER) show a considerable improvement using our novel amendments over the method of [Gough & Way, 2003]. These results are confirmed in the manual evaluation on a 200-sentence subset obtained at random from the larger testset: with respect to intelligibility, we outperform *Logomedia* by 3.5%, and by 18.5% when accuracy is measured. These too are considerable improvements on the figures reported in [Gough & Way, 2003].

### 4.2 English-French: Controlling the Source Language

#### 4.2.1 Automatic Evaluation

We also develop an English-French EBMT system trained on the same data, and using the same techniques. As far as we are aware, this is the first research which attempts to filter the source language data using controlled language specifications in an EBMT system.

**Table 5.** Summary of results in comparing our revised EBMT (English-French) system with *Logomedia* and [Gough & Way, 2003] using the Automatic Evaluation Metrics on a 3885 Sentence Testset

| Experiment | Precision | Recall | Bleu | WER | SER |
|---|---|---|---|---|---|
| Alignment 1 [Gough/Way 03] | 0.3081 | 0.4477 | 0.0925 | 71.8 | 93 |
| Alignment 2 | 0.3115 | 0.4566 | 0.0954 | 70.0 | 92 |
| Top 10% words corrected | 0.3216 | 0.4756 | 0.1016 | 68.5 | 90 |
| Top 30% words corrected | 0.3551 | 0.4880 | 0.1147 | 67.1 | 89 |
| Additional word Translations | 0.3891 | 0.5293 | 0.1352 | 64.8 | 84 |
| Logomedia | 0.3554 | 0.3724 | 0.2321 | 64.7 | 90.2 |

What is notable about the results in Table 5 is that they paint a somewhat confusing picture: while Bleu shows that *Logomedia* outperforms our system by quite a margin, the Precision and Recall figures show precisely the opposite. We comment further on this in section 4.3.

#### 4.2.2 Manual Evaluation

Given the contradictory nature of the results obtained in the automatic evaluation, we carried out a manual evaluation using the same 200-sentence testset, and the same metrics of intelligibility and accuracy using the same scale as before.

Overall *Logomedia* outperforms our system with respect to intelligibility. For scores 2, 3 and exact match, a total of 188 (94%) translations are counted

for our system, while for *Logomedia* this figure is higher at 195 (97.5%). As far as accuracy is concerned, however, 80% of translations produced for *Logomedia* achieve a score of 3,4 or better, while we score 3 or above in 90% of cases.

#### 4.2.3 Summary

Interestingly, the Bleu scores show that our approach is about 26% less successful in translating in this direction than for French-English. In addition, they show that *Logomedia* outperforms our system. Interestingly, *Logomedia* does about 4.2% better for English-French than it does in the other direction using the same Bleu indicator.

However, in obtaining Precision and Recall figures, we observe that contrary to the Bleu scores, our system not only significantly improves in the direction English-French (Precision 39%, Recall 53%: French-English: Precision 30%, Recall 36%), but also that we outperform *Logomedia* (Precision 35.5%, Recall 37%), especially for Recall. Figures calculated for WER and SER also suggest that translations from English-French are better than those generated from French-English. We comment further on these results in the next section.

The results from the manual evaluation also appear to show that our system performs better in the direction English-French than for French-English: as regards intelligibility, 188/200 (94%) translations score 2 or higher for English-French, while 184 (92%) score at least 2 in the other direction; for accuracy, 180 (90%) translations score 3, 4 or better for English-French, while 174 (87%) score at least 3 for French-English. While these results are quite close, they appear to side with the Precision and Recall results—indicating that controlling the source language produces better results—and providing some evidence that the Bleu scores may be anomalous.

### 4.3 Evaluating Evaluation Metrics

Automated metrics such as Bleu enable MT developers to evaluate potentially huge amounts of data without any human intervention. As an example, note that while the *wEBMT* system evaluated in [Way & Gough, 2003] used a testset of just 200 translations, the research presented in [Gough & Way, 2003] and here evaluates 3885 translations. As such, the benefits of such evaluation measures cannot be overlooked.

Nevertheless, given the requirements of conference organisers and scientific programme committees that we use and publish detailed evaluations using metrics such as Bleu, we must be sure that what we are using are useful and accurate measures. Furthermore, one of the main reasons such metrics were introduced was to try to overcome the high costs of conducting human evaluations. In that regard, we must be certain that automatic evaluation techniques correlate accurately with human judgements.

In the field of MT, the issue of automatic evaluation metrics is currently a hot topic: recall the panel session on the 'Holy Grail', together with a number of other papers which focused on MT evaluation metrics at the recent MT Summit. [Turian *et al.,* 2003] find the F-measure to be a more reliable metric than the Bleu and NIST measures, while [Coughlin, 2003] emphasises the preference of human evaluation but proposes Bleu and NIST as reliable alternatives.

Our results show that different such metrics demonstrate conflicting results—surely if such metrics are to be at all objective, they should deliver similar results. The figures for Precision and Recall obtained for our system suggest that controlling the source and translating from English to French produces better translations. The figures obtained for Word Error Rate and Sentence Error Rate confirm this and the results receive further corroboration via a manual evaluation of 200 sentences using the traditional metrics of Accuracy and Intelligibility. However, the Bleu scores produced for the same data imply that translating from French-English (controlled generation) generates better translations. In that regard, our study corroborates the findings of [Turian *et al*., 2003] that traditional NLP measures such as Precision and Recall are more reliable than Bleu. Finally, these conflicting results show that the current debate as to the relative merits of automatic evaluation metrics will no doubt continue—using standard measures is a good thing, but none of us want poor, unreliable metrics to become the norm.

## 5. Concluding Remarks

In this paper we have presented an EBMT system where, as in [Gough & Way, 2003], the generation of the target string is filtered by data written according to controlled language specifications. Given the same data as [Gough & Way, 2003], we applied an improved sub-sentential alignment algorithm to automatically extract additional lexical resources. We consider that the research reported here is encouraging, in that it shows that an EBMT

system can outperform a good, on-line MT system such as *Logomedia* using automatic evaluation metrics. Given that the results reported in [Way & Gough, 2003] on uncontrolled data showed the reverse to be true, we consider that our work tends to confirm the hypothesis of [Schäler et al, 2003; Carl, 2003] that EBMT systems ought to outperform rule-based systems when confronted with data written according to controlled language specifications.

With respect to controlled analysis, we have presented an English-French EBMT system trained on the same data as [Gough & Way, 2003]. As far as we are aware, this is the first research which attempts to filter the source language data using controlled specifications in an EBMT system. We compare the results obtained with those produced from French-English, by carrying out both automated and manual evaluations. The figures for manual evaluation, Precision and Recall and WER/SER suggest that our system produces better translations in the direction of English-French. It may be, therefore, that controlling the source text is generally more effective than attempting to control the output translations. However, the Bleu scores show a preference for the French-English translations, i.e. controlled synthesis. Nevertheless, the Bleu scores are not in line with the other evaluation metrics, which leads us to agree with [Turian *et al*., 2003] that Precision and Recall may be more reliable metrics. These conflicting results highlight the need for further assessment of the reliability of automatic evaluation metrics in MT.

## References


[1] Katy Barthe. 1998. GIFAS Rationalised French: Designing one Controlled Language to Match another. In *CLAW*98: *Proceedings of the Second International Workshop on Controlled Language Applications*, Pittsburgh, PA., pp.87—102.

[2] Arendse Bernth. 2003. Controlled Generation for Speech-to-Speech MT Systems. In *EAMT-CLAW 03*, *Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Proceedings*, Dublin, Ireland, pp.1—7.

[3] Michael Carl. 2003. Data-Assisted Controlled Translation. In *EAMT-CLAW 03*, *Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Proceedings*, Dublin, Ireland, pp.16—24.

[4] Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. In *MT Summit IX,* New Orleans, LA., pp.63—70.

[5] Nano Gough and Andy Way. 2003. Controlled Generation in Example-Based Machine Translation. In *MT Summit IX,* New Orleans, LA., pp.133—140.

[6] Nano Gough, Andy Way and Mary Hearne. 2002. Example-Based Machine Translation via the Web. In S. Richardson (ed.) *Machine Translation: From Research to Real Users. 5th Conference of the Association for Machine Translation in the Americas (AMTA-2002)*, LNAI 2499, Springer Verlag, Berlin/Heidelberg, Germany, pp.74—83.

[7] Thomas R.G. Green. 1979. The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior* **18**:481—496.

[8] Anthony Hartley, Donia Scott, John Bateman and Danail Dochev. 2001. AGILE – A System for Multilingual Generation of Technical Instructions. In *MT Summit VIII, Machine Translation in the Information Age, Proceedings*, B. Maegaard. Ed., Santiago de Compostela, Spain, pp.145—150.

[9] Patrick Juola. 1994. A Psycholinguistic Approach to Corpus-Based Machine Translation. In *CSNLP 1994; 3rd International Conference on the Cognitive Science of Natural Language Processing*, Dublin, Ireland, [pages not numbered].

[10] Linda Means and Kurt Godden. 1996. The Controlled Automotive Service Language (CASL) Project. In *CLAW 96*: *Proceedings of the First International Workshop on Controlled Language Applications*, Leuven, Belgium, pp.106—114.

[11] Teruko Mitamura and Eric Nyberg. 1995. Controlled English for Knowledge Based MT: Experience with the KANT System. In *Proceedings of Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, pp.158—172.

[12] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the ACL, Philadelphia PA.,* pp.311—318.

[13] Richard Power, Donia Scott and Anthony Hartley. 2003. Multilingual Generation of Controlled Languages. In *EAMT-CLAW 03, Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation, Proceedings*, Dublin, Ireland, pp.115—123.

[14] Reinhard Schäler, Andy Way and Michael Carl. 2003. Example-Based Machine Translation in a Controlled Environment. In M. Carl & A. Way (eds.) *Recent Advances in Example-Based Machine Translation*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.83—114.

[15] Joseph Turian, Luke Shen and Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *MT Summit IX,* New Orleans, LA., pp.386—393.

[16] Tony Veale and Andy Way. 1997. *Gaijin:* A Bootstrapping, Template-Driven Approach to Example-Based Machine Translation. In *International Conference, Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria, pp.239—244.

[17] Andy Way and Nano Gough. 2003. *wEBMT*: Developing and Validating an Example-Based Machine Translation System using the World Wide Web. *Computational Linguistics* **29**(3), forthcoming.

[18] Setsuo Yamada, Eiichiro Sumita and Hideki Kashioka. 2000. Translation Using Information on Dialogue Participants. In *Proceedings of the 6th Applied Natural Language Conference and 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA., pp.37—43.