Example-Based Machine Translation using the Marker Hypothesis

Nano Gough

Bachelor of Science in Applied Computational Linguistics

A dissertation submitted in fulfilment of the

requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

DCU

Dublin City University

School of School of Computing

Supervisor: Dr. Andy Way

January, 2005

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed Mono Gough

Student ID 97024635

Date January, 2005

Contents

Abstract v

1 Introduction 1

1.1 An Example-Based Approach 2

1.2 A Linguistics-Lite Approach 3

1.3 Experiments and Contributions 5

1.3.1 Phrase-Based EBMT 6

1.3.2 Controlled EBMT 8

1.3.3 Scalable EBMT 9

1.4 Structure of Thesis 10

2 Corpus-Based Machine Translation: Background and Motivation 12

2.1 CBMT versus RBMT 13

2.1.1 RBMT 14

2.1.2 Advantages of CBMT 16

2.2 What is CBMT? 18

2.2.1 CBMT versus Translation Memory 18

2.2.2 SMT vs EBMT 20

2.3 What is EBMT? 23

2.4 Related Research 26

2.4.1 Word-Based Matching 26

2.4.2 Applying Dynamic Programming Techniques 28

2.4.3 Applying CBR techniques 32

2.4.4 Generalised Templates 34

2.4.5 Structural Approaches 40

2.4.6 Recombination and Boundary Friction 44

2.5 Marker-Based EBMT 47

2.5.1 The Marker Hypothesis as a Non-Declarative Grammar 51

2.6 Summary 53

3 Phrase-Based EBMT 55

3.1 Motivation for a Phrase-Based Model 55

3.2 Seeding the Example-Base 59

3.2.1 The Phrasal-Lexicon 59

3.2.2 The Marker-Lexicon 60

3.2.3 The Word-level Lexicon 63

3.2.4 The Generalized lexicon 63

3.2.5 Summary of Knowledge Sources 64

3.3 Retrieving chunks and Performing Translation 65

3.4 Experiments and Results 69

3.4.1 Experiments using Single Knowledge Sources: Sentences 70

3.4.2 Experiments using Single Knowledge Sources: Noun Phrases 74

3.4.3 Experiments using Multiple Knowledge Sources 75

3.4.4 Producing Weights and Ranking Translations 78

3.4.5 Relative Gain of EBMT 84

3.4.6 Evaluating individual on-line MT systems 86

3.5 Web-based *post hoc* Validation 88

3.5.1 Determiner-Noun Agreement 89

3.5.2 Noun-Verb Agreement 91

3.6 Discussion 94

4 Controlled EBMT 97

4.1 Controlled Translation 98

4.2 Our Controlled EBMT System 100

4.2.1 An Improved Sub-Sentential Alignment Algorithm 103

4.2.2 Alignment Evaluation 110

4.3 Automatic Evaluation Metrics . . . 113
4.3.1 SER and WER . . . 114
4.3.2 The BLEU metric . . . 115
4.3.3 Precision and Recall . . . 116
4.4 Experiments and Results . . . 119
4.4.1 Controlling the Target Language: French-English . . . 121
4.4.2 Controlling the Source Language: English-French . . . 125
4.4.3 Controlling the Source Language versus Controlling the Target Language . . . 127
4.4.4 Discussion of Evaluation Metrics . . . 128
4.5 *Post Hoc* Validation . . . 133
4.6 Ranking Translations . . . 134
4.7 Discussion . . . 136
5 Scalability **139**
5.1 Scalability in EBMT . . . 140
5.2 Scalability in Marker-Based EBMT . . . 142
5.3 Experiments and Results: An Automatic Evaluation . . . 144
5.3.1 Evaluation of the Complete Test Set . . . 145
5.3.2 Evaluation of Non-Exact Matches . . . 148
5.3.3 Filtering the Data . . . 149
5.3.4 Integrating Mutual Information . . . 151
5.4 Experiments and Results: A Manual Evaluation . . . 156
5.4.1 Evaluating English-French Translations . . . 156
5.4.2 Evaluating French-English Translations . . . 157
5.5 Discussion of Results . . . 158
5.6 Comparing our marker-based EBMT system to an SMT system . . . 161
5.7 Discussion . . . 165
6 Conclusions **167**
6.1 Contributions of this Thesis . . . 169

6.1.1 Phrase-Based EBMT 169

6.1.2 Controlled EBMT 170

6.1.3 Scalable EBMT 171

6.2 Future Research Avenues 171

6.2.1 Alternative Bitexts 171

6.2.2 Other Language Pairs 172

6.2.3 Recombination 174

6.2.4 Efficiency 176

6.3 Closing Remarks 176

Bibliography **178**

Abstract

The development of large-scale rules and grammars for a Rule-Based Machine Translation (RBMT) system is labour-intensive, error-prone and expensive. Current research in Machine Translation (MT) tends to focus on the development of corpus-based systems which can overcome the problem of knowledge acquisition.

Corpus-Based Machine Translation (CBMT) can take the form of Statistical Machine Translation (SMT) or Example-Based Machine Translation (EBMT). Despite the benefits of EBMT, SMT is currently the dominant paradigm and many systems classified as example-based integrate additional rule-based and statistical techniques. The benefits of an EBMT system which does not require extensive linguistic resources and can produce reasonably intelligible and accurate translations cannot be overlooked. We show that our *linguistics-lite* EBMT system can outperform an SMT system trained on the same data.

The work reported in this thesis describes the development of a *linguistics-lite* EBMT system which does not have recourse to extensive linguistic resources. We apply the Marker Hypothesis (Green, 1979) — a psycholinguistic theory which states that all natural languages are 'marked' for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes. We use this technique in different environments to segment aligned ⟨English, French⟩ phrases and sentences. We then apply an alignment algorithm which can deduce smaller aligned chunks and words. Following a process similar to (Block, 2000), we generalise these alignments by replacing certain function words with an associated tag. In so doing, we cluster on marker words and add flexibility to our matching process. In a *post hoc* stage we treat the World Wide Web as a large corpus and validate and correct instances of determiner-noun and noun-verb boundary friction.

We have applied our marker-based EBMT system to different bitexts and have explored its applicability in various environments. We have developed a phrase-based EBMT system (Gough et al., 2002; Way and Gough, 2003). We show that despite the perceived low quality of on-line MT systems, our EBMT system can produce good quality translations when such systems are used to seed its memories.

(Carl, 2003a; Schäler et al., 2003) suggest that EBMT is more suited to controlled translation than RBMT as it has been known to overcome the 'knowledge acquisition bottleneck'. To this end, we developed the first controlled EBMT system (Gough and Way, 2003; Way and Gough, 2004). Given the lack of controlled bitexts, we used an on-line MT system *Logomedia* to translate a set of controlled English sentences. We performed experiments using controlled analysis and generation and assessed the performance of our system at each stage. We made a number of improvements to our sub-sentential alignment algorithm and following some minimal adjustments to our system, we show that our controlled EBMT system can outperform an RBMT system.

We applied the Marker Hypothesis to a more scalable data set. We trained our system on 203,529 sentences extracted from a *Sun Microsystems* Translation Memory. We thus reduced problems of data-sparseness and limited our dependence on *Logomedia*. We show that scaling up data in a marker-based EBMT system improves the quality of our translations. We also report on the benefits of extracting lexical equivalences from the corpus using Mutual Information.

Acknowledgements

There are a number of people who have contributed in no small way to the completion of this thesis. Firstly, I would like to thank Andy Way. He has been an excellent supervisor, always approachable and full of support, encouragement and reassurance.

Thanks to Aoife, Mary and Michelle who have constantly been there to provide support, advice and friendship. Thanks to all of you (especially Mary) for always being willing to discuss ideas and lend a hand when problems arose.

Thanks to Declan for all his help and insightful comments and to Mick, Ruth, Sara and Regina for providing a pleasant working environment and for generally putting up with me while my thesis-writing took over. Thanks also to Cathal for making me realise that writing a thesis is possible, and that all my hard work would pay off in the end.

I would like to thank all of my friends outside of DCU. It is impossible to mention everyone but special thanks must go to Caroline and Anne-Marie who have been brilliant housemates and friends. Thanks for helping me through the tough times and taking my mind off my thesis every now and again. Thanks also to Niamh and Elaine who have been the best of friends and have never complained when my thesis took priority over almost everything. Thanks to Maura and Ethna for driving me and my laptop around from time to time and for making things that bit easier for me. Thanks to all the friends I made in Boston for reminding me of life outside of thesis-writing.

Finally, I would like to especially thank my parents, Philip and Marianne for always believing in me and for their endless support, love and encouragement. To my brothers, Philip and Pádraig for giving me space to write my thesis and providing a constant source of entertainment. To Clara, my sister and best friend, who probably knows more about Machine Translation at this stage than she could ever have imagined! Thanks for being a great sister and for always being there whenever you were needed, no questions asked.

I am grateful to both IBM and DCU for their financial support of the work carried out in this thesis. Thanks also to Sun Microsystems for providing the resources necessary to carry out certain aspects of this research.

Chapter 1

Introduction

Nowadays, the objective of Machine Translation (MT) is not the generation of a consistently 'perfect' translation but rather the production of useful systems which can help reduce the amount of manual translation involved and speed up the work of the human translator. A system which can be developed quickly without requiring high-level linguistic expertise can potentially be very useful if the translations produced are of a reasonable standard and can be interpreted by the user.

Recent years have seen a move away from rule-based approaches to MT and the benefits associated with corpus-based approaches have come to light. Traditional, rule-based approaches to MT require the development of large-scale grammars and rules. This generally requires linguistic expertise and, as such, a substantial amount of manual labour. This problem has come to be known as 'the knowledge acquisition bottleneck'.

Recent research (Brown, 2003; Cicekli and Güvenir, 2003; Xia and McCord, 2004) has focused on automatically inferring rules — similar to those used in a rule-based system — from corpora. This could spark the revival of rule-based methods in MT. Indeed, an emerging viewpoint (Carl et al., 2002; Langlais and Simard, 2002) seems to favour the development of hybrid systems — the integration of corpus-based and rule-based techniques often in a sublanguage domain. This is possibly the way forward for the development of high quality, sophisticated commercial MT systems. Before such a goal can be achieved, however, the optimal development of the various approaches to MT is necessary.

1.1 An Example-Based Approach

Corpus-based Machine Translation (CBMT) requires an aligned bilingual corpus as a prerequisite. Novel translations are derived by extracting and recombining relevant fragments from these existing translations. Corpus-based systems are not generally associated with the manual development of rules and grammars and thus can overcome the 'knowledge acquisition bottleneck' that Rule-Based Machine Translation (RBMT) systems are prone to. Accordingly, recent research in MT has drawn on corpus-based rather than rule-based techniques.

CBMT includes two approaches — Example-Based Machine Translation (EBMT) and Statistical Machine Translation (SMT) (cf. section 2.2). The latter generally requires large-scale corpora from which a probability model is derived. Translations are produced by maximising the statistical probabilities, i.e. the probability that a target string t occurs and the probability that a source string s translates as t. EBMT does not generally integrate a complex probability model. Instead, new input i is matched against one or more existing source examples s and the relevant target fragments are extracted and recombined to derive a translation, t.

As we have already pointed out, most current research in MT tends to focus on corpus-based methods. Within this domain, SMT seems to be the dominant approach. The earliest SMT systems were modelled on the approach initiated by (Brown et al., 1988) and integrated only word-level correspondences. More recent research in SMT (Och et al., 1999; Yamada and Knight, 2001; Marcu and Wong, 2002; Charniak et al., 2003; Koehn et al., 2003) has realised the potential of including phrasal alignments. Furthermore, the intuition that using syntax-based models should improve the performance of SMT has been exploited in some recently developed systems (Yamada and Knight, 2001; Charniak et al., 2003; Aue et al., 2004). Although these techniques are relatively novel additions to SMT, the results are promising.

However, the idea of storing phrasal or chunk correspondences is not new to EBMT. Moreover, example-based systems integrate at least some level of syntactic information, although to varying degrees. That is, EBMT has always employed techniques which only recently have been applied to SMT with positive results. Furthermore, given that EBMT systems do not rely on probabilities derived from large-scale corpora, they can be developed using much smaller corpus resources than are required for purely statistical models.

Despite this, SMT remains the favoured approach. Moreover, perhaps due to the potential of the hybrid model, systems classified as example-based regularly borrow from rule-based and statistical techniques. It is possible, therefore, that the benefits of EBMT are not fully realised or exploited in current approaches.

This poses a research question: How useful is an EBMT system which does not integrate high-level linguistic information? Furthermore, can such an approach be shown to outperform rule-based and statistical approaches? To this end, we propose the development of an EBMT system which does not have recourse to extensive linguistic resources.

1.2 A Linguistics-Lite Approach

Some EBMT systems use high-level linguistic information and can be described as *linguistics-rich* systems. Others do not integrate such extensive linguistic knowledge and can be regarded as *linguistics-lite.* 'Purist' approaches to EBMT which typically do not integrate any additional linguistic resources aside from the given corpus have not been widely explored by researchers. Somers et al. (1994) attempted such an approach but, unsurprisingly, found they were limited by the lack of linguistic information. Those systems which store examples as annotated tree structures are by their nature linguistically rich approaches (Hearne and Way, 2003; Way, 2003). These types of systems have the disadvantage of requiring extensive computational and linguistic resources which are currently hard to come by, especially for certain language pairs.

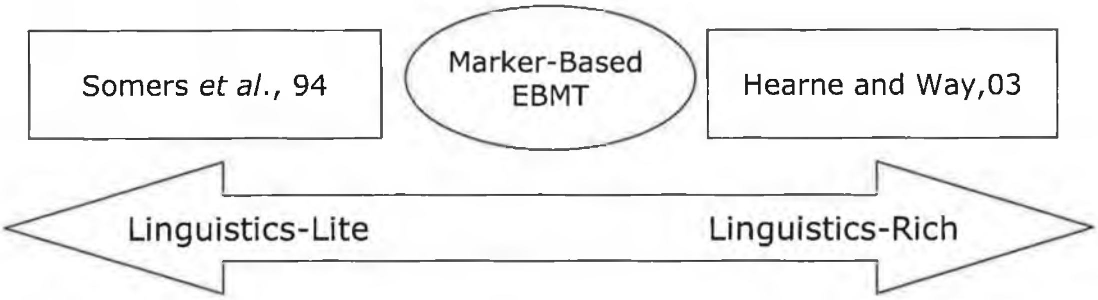


Figure 1.1: Linguistics-Lite and Linguistics-Rich approaches to EBMT

The approach to EBMT which is presented in this thesis can be categorised as *linguistics-lite* and lies somewhere between the approaches of (Somers et al., 1994) and (Hearne and Way, 2003) as shown in Figure 1.1. Our goal is to extend the system as far as possible without recourse to high-level linguistic techniques. The only linguistic information which we apply is that of the Marker Hypothesis (Green, 1979). The Marker Hypothesis (cf. section 3.2.2) is a theory rooted in psycholinguistics which states that all natural languages are 'marked' for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes.

The Marker Hypothesis has previously been exploited for the purpose of MT in the *Gaijin* system (Veale and Way, 1997) using ⟨English, German⟩ corpora and in the *METLA* system by (Juola, 1994, 1997) in relation to ⟨English, French⟩ and ⟨English, Urdu⟩ (cf. section 2.5).

The methodology behind EBMT involves matching new input against existing examples and extracting and combining relevant fragments to produce new translations. As the chances of finding a match against smaller examples are higher, most EBMT systems realise the benefits of extending correspondences to a sub-sentential level.

In some cases, these correspondences are extracted dynamically at run-time and there is little or no pre-processing of the corpus. Each time the system runs, the correspondences have to be re-computed and are not stored in the system's memories for future use (Sumita, 2003). Other approaches extract generalised translation templates or patterns from the corpus in a pre-processing stage for use in the matching process (Kaji et al., 1992; Cicekli and Güvenir, 2003). Generalisations are very useful in reducing the amount of data required and for providing a broader, more general template for matching. Structural or derivational approaches (Hearne and Way, 2003; Watanabe et al., 2003; Way, 2003) store examples as annotated trees, produced by parsing techniques. These structures are explicitly linked at the level of words and/or phrases. Different approaches make use of various linguistic resources, such as parsers, morphological analysers, taggers, dictionaries and thesauri.

In section 2.4, we describe various approaches to EBMT. Our *linguistics-lite* methodology integrates several features applied in these different systems. As in (Veale and Way, 1997), we apply the Marker Hypothesis to segment ⟨source, target⟩ examples according to a closed set of marker words and subsequently extract sub-sententially-aligned chunks and words from the corpus in a pre-processing stage. These are used to seed a marker-lexicon and a word-level lexicon in our system's memories.

Initially, we apply a näive, yet effective sub-sentential alignment algorithm, which aligns source and target chunks on condition that their marker tags matched sequentially. Following the revision of this algorithm, we integrate a bilingual dictionary to provide us with an initial set of ⟨source, target⟩ word correspondences to aid the sub-sentential alignment process. However, we later derive lexical equivalences from the corpus itself. Consequently, we reduce our dependency on the dictionary and improve the quality of the chunk alignments generated. We also significantly increase the number of entries in our word-level lexicon.

In addition to storing a set of sub-sententially-aligned strings, we create a set of generalised templates. Given the *linguistics-lite* nature of our approach, we draw our methodology from the work of (Block, 2000) (cf. section 2.4.4), where little additional knowledge aside from the corpus is required. In contrast to Block's method we do not apply a statistical word alignment tool, but instead use a set of marker tags to identify words which can be replaced by variables. Therefore, without using extensive linguistic resources or complex parsing techniques, we derive a set of additional aligned fragments in the form of words, phrases and templates from a bitext. These lexical resources are used to seed the memories of our EBMT system. Matching can, therefore, occur against strings or generalised templates. Figure 1.2 shows the different lexical resources used to seed the memories of our EBMT system.

We will show how despite integrating 'low-level' linguistic information, our system can produce good quality translations and can be seen to outperform both statistical and rule-based MT.

1.3 Experiments and Contributions

In this thesis we explore the development of an ⟨English, French⟩ EBMT system based on the Marker Hypothesis. We perform our experiments in three separate paradigms, using

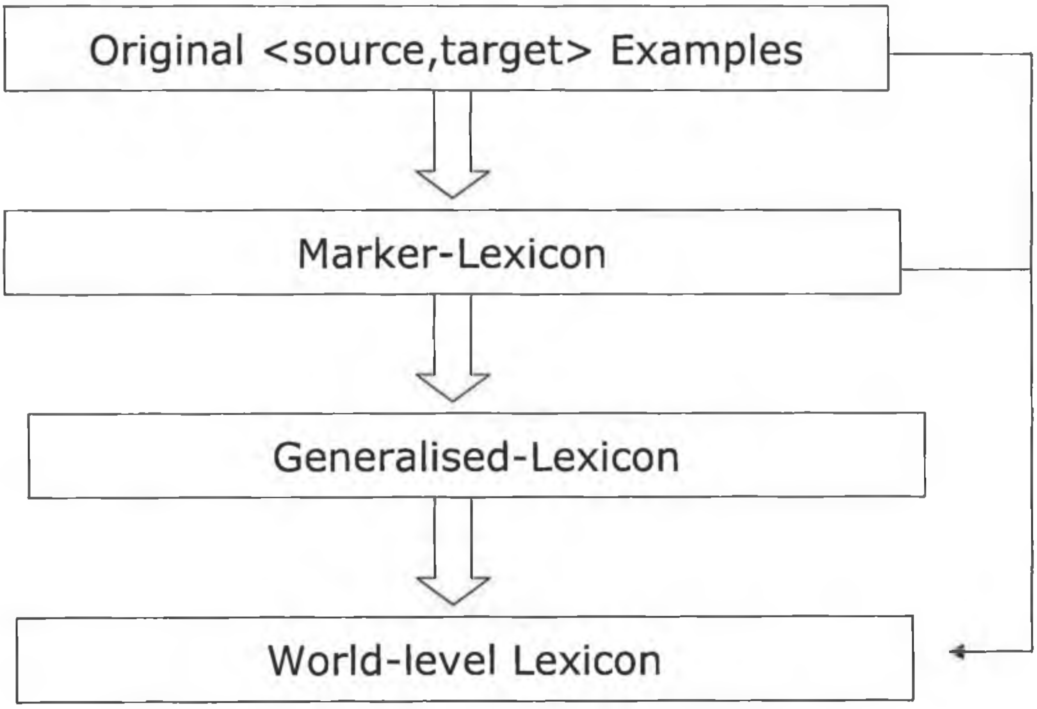


Figure 1.2: Resources used to seed our example-base: the marker-lexicon, generalised-lexicon and word-level lexicon are derived via the Marker Hypothesis

the following bitexts to seed the memories of our system:

- 218,697 English phrases from the Penn-II Treebank and their French translations derived from three on-line RBMT systems;
- 1,691 English sentences written according to *Sun Microsystems* controlled language specifications and their French translations derived from one on-line RBMT system;
- 203,529 ⟨English, French⟩ sententially-aligned strings extracted from a *Sun Microsystems* TM.

1.3.1 Phrase-Based EBMT

Recent research (Carl et al., 2002; Schäler et al., 2003) has highlighted the potential of Translation Memory (TM) technology. Comparisons can be drawn between TMs and EBMT systems. Both EBMT and TM match new input against existing examples. However, this is where the similarity ends. TMs operate as computer-assisted translation

(CAT) tools. Given an input sentence, fuzzy matching techniques return a set of 'close' matches and the translator extracts and recombines relevant target fragments to produce a translation. TMs, therefore, do not perform translation. This can be contrasted with an EBMT system where the stages of extracting relevant fragments and recombining these to produce a translation are fully automated.

In a TM, matching is performed against a set of sententially-aligned pairs. Given that the chance of finding an exact match increases when matching is performed against smaller fragments, current TM technology is under-exploited. Most EBMT systems integrate subsentential alignments. (Carl et al., 2002; Schäler et al., 2003) propose that TMs can be developed into MT engines via the integration of the phrasal-lexicon (PL) in EBMT. These novel ideas prompted us to develop a phrase-based EBMT system.

In an initial experiment, given that we did not have recourse to a set of sententially-aligned examples, we adopted a novel approach where we used on-line RBMT systems to seed our system's memories (Gough et al., 2002; Way and Gough, 2003). We created a phrase-based EBMT system by extracting English phrases from the Penn-II Treebank and automatically deriving their French translations via three different on-line MT systems, namely:

- SDL International's Enterprise Translation Server (system A)[1];
- Reverso by Softissimo (system B)[2];
- Logomedia (system C)[3].

We then applied the Marker Hypothesis as in (Veale and Way, 1997) using a naïve alignment algorithm to deduce additional resources, i.e., smaller aligned chunks, words and generalised templates.

Numerous experiments were performed to analyse the effects of using various combinations of the translations produced by the on-line systems to seed our example-base. We performed a manual evaluation of the translations produced. Despite the fact that on-line systems are perceived to be of low quality, we show that by using the systems listed above

[1] http://www.freetranslation.com

[2] http://www.reverso.net/text_translation.asp

[3] http://www.logomedia.net

to derive our example-base and applying a naïve sub-sentential alignment algorithm, our system produces good quality translations.

In a manual evaluation, we demonstrate how increasing the number of fragments used to seed our example-base improves the quality of translations produced. Like many CBMT systems, we can output numerous candidate translations. We assign these translations a weight and rank them accordingly. We show how the 'best' translation can be consistently located within the top 1% of translation candidates generated by the EBMT system, thus facilitating any *post hoc* human interaction with the system. As a spin-off of the experiments and evaluation performed, we can identify the 'best' on-line RBMT system among those used to seed our example-base. We also show how our EBMT system can outperform the rule-based systems.

EBMT systems are prone to problems of boundary friction (cf. section 2.4.6). This is a consequence of stitching together chunks and words which have been extracted from different contexts, causing the resulting translation to be ungrammatical. We apply an approach similar to that of (Grefenstette, 1999) where we treat the World Wide Web (WWW) as a large corpus and use it to correct and validate the translations produced by our system so as to reduce problems of boundary friction.

1.3.2 Controlled EBMT

(Carl, 2003b; Schäler et al., 2003) propose that EBMT is more applicable to controlled language (CL) translation than RBMT. They point out that the difficulty in performing controlled RBMT, is that at each stage in an RBMT system, i.e. analysis, transfer and generation, a level of control must be exerted if the system is to produce a controlled translation which is of high quality. EBMT systems, they suggest, are more suited to controlled translation as they have been shown to avoid the knowledge acquisition bottleneck associated with RBMT. Moreover, given that the quality of translations generated via EBMT is largely dependent on the quality of the examples, exerting control over these examples should positively affect the resulting translations.

In a second experiment we developed the first controlled EBMT system (Gough and Way, 2003; Way and Gough, 2004). In this experiment, we seeded our example-base with sentences rather than phrases. Given the lack of controlled bitexts, we translated a set of controlled English sentences into French via the on-line system *Logomedia*. As in our phrase-based system, we applied the Marker Hypothesis to derive a set of sub-sentential alignments and templates to seed our system's memories. By translating a set of uncontrolled sentences from English-French and French-English, we performed the first research on filtering the source and target languages using controlled data specifications in an EBMT system.

We integrate automatic evaluation metrics and, as such, we are able to perform a more extensive evaluation than in our phrase-based system. Furthermore, we provide some analysis of these metrics and discuss their reliability and consistency with a human evaluation. We show how improving our sub-sentential alignment algorithm and minimally adjusting our lexical resources leads to an improved translation performance. We provide a baseline comparison with the on-line RBMT system *Logomedia* and show how we can outperform the rule-based system. Consequently, we show that we can support the claims of (Carl, 2003a; Schäler et al., 2003) that EBMT is probably more suitable for controlled translation than RBMT.

1.3.3 Scalable EBMT

In order to assess the effects of using much larger training data to seed our system's memories, we scaled up our training data by using a *Sun Microsystems* TM to derive our lexical resources and seed the memories of our EBMT system (Gough and Way, 2004). In this way, we reduced the problems of data-sparseness present in our controlled EBMT system and significantly reduced our dependency on the on-line system *Logomedia* from previous marker-based models.

Confining translation to a particular sublanguage domain means that the number of words and phrases for translation can potentially be reduced to a smaller subset of the language in question. Moreover, in a restricted domain, elements are more likely to occur quite commonly, thereby adding reliability to the weighting measures applied in our system. To this end, we increased the similarity between our training data and our test set and confined translation to the domain of computer manuals.

We performed several experiments to test the quality of our translations, using both manual and automatic evaluation metrics. We integrated a novel filtering technique and assessed what effects this had on translation quality. We made further improvements to our alignment algorithm by applying Mutual Information (MI) to derive lexical equivalences from our bitext. As a result, we also significantly increased the size of our word-level lexicon. In a comparison with those systems listed in (Somers, 2003), this system is the largest English-French EBMT system. We show that using large-scale data and integrating MI also improves translation performance in marker-based EBMT.

Finally, we compared our *linguistics-lite* EBMT system with an SMT system. We used *Giza++*(Och and Ney, 2003)[4], in conjunction with the CMU-Cambridge statistical toolkit[5] and and the ISI ReWrite Decoder[6] to derive a probability model for an SMT system which integrates word correspondences. We trained this SMT system and our EBMT system on the same data and show that even without the integration of high-level linguistic information, EBMT can outperform SMT.

1.4 Structure of Thesis

This thesis is structured as follows:

Chapter 2 describes corpus-based and rule-based approaches to MT and explains our motivation for research in the EBMT paradigm. This chapter also describes different approaches to EBMT and the techniques and methodologies applied by different systems. In addition, it gives details of the Marker Hypothesis and where it has been applied previously in Natural Language Processing (NLP) applications and specifically in MT.

Chapter 3 describes our experiment with phrase-based EBMT. Translation is performed from English to French and a detailed manual evaluation is carried out to assess the quality and coverage of the translations produced. As a by-product of our research, we evaluate the rule-based systems used to seed our example-base and compare the

[4] http://www.isi.edu/~och/Giza++.html

[5] http://mi.eng.cam.ac.uk/ prc14/toolkit.html

[6] http://www.isi.edu/licensed-sw/rewrite-decoder/

performance of EBMT with the individual performances of these on-line rule-based systems. Our *post hoc* validation method is also described in detail.

Chapter 4 describes the development of a controlled EBMT system. Two experiments are presented. In the first, the analysis stage is controlled and in the second, control is exerted at the generation stage. Detailed automatic and manual evaluations are carried out on the translations produced and a comparison is provided with the on-line system *Logomedia*. Improvements made to the sub-sentential alignment algorithm are also outlined.

Chapter 5 describes how using a much larger amount of training data improved translation performance in a marker-based EBMT system. We report on the integration of MI in our scalable model and outline an experiment to compare our EBMT system with a statistical system.

Chapter 6 concludes and provides a summary and discussion of the novel research presented in this thesis. Potential avenues for future work are also presented.

Güvenir, 2003; Xia and McCord, 2004). Although such approaches are often classified as corpus-based, the rules inferred can be largely similar to those applied in rule-based systems. It seems, therefore, that the distinction between corpus-based and rule-based approaches is becoming less marked. It is reasonable to propose, therefore, that the 'knowledge acquisition bottleneck' which stems from the requirement of extensive lexical and grammatical resources might be overcome and that RBMT may yet flourish, albeit within a hybrid environment.

In this chapter we will outline the advantages of corpus-based approaches and our motivation for research in EBMT. We will describe related research in EBMT and provide details of the Marker Hypothesis and its role in previous NLP and specifically, MT applications.

2.1 CBMT versus RBMT

Currently, there is a growing awareness that corpus resources can be exploited for different NLP tasks, and MT is no exception to this trend. The increasing number of bilingual corpora available and the rapid expansion of the WWW has encouraged research in MT in a different, potentially more positive direction. Data-driven, empirical techniques which exploit corpora to produce translations, can be classified as SMT and EBMT.

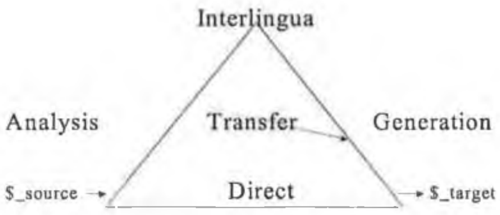


Figure 2.1: The 'Vauquois pyramid'

Chapter 2

Corpus-Based Machine Translation: Background and Motivation

Over the past decade or so, research in MT has evolved and expanded into different paradigms. Traditionally, the translation of natural languages by computer was implemented via dictionaries, large-scale grammars and rules. A major turning point however, came in the mid-to-late 1980's with the advent of corpus-based methods.

The concept of EBMT stems from a paper presented by Makoto Nagao (Nagao, 1984). Nagao described a process of 'machine translation by analogy', which suggested the use of a bilingual corpus as a resource for producing new translations.

While Nagao's ideas were initially met with a degree of skepticism, the limitations associated with traditional RBMT systems were coming to light and the time was ripe for fresh ideas. When IBM's Peter Brown introduced an entirely new and 'purely statistical' approach at the second TMI conference in 1988, a wave of research using corpora as opposed to traditional rules ensued. There followed a period where the 'old' and 'new' approaches were viewed as conflicting. In the current climate, however, the trend is towards CBMT rather than RBMT.

Having said that, it is worth referring to work which involves the automatic extraction of patterns or rules from corpora (Kaji et al., 1992; Watanabe et al., 2000; Cicekli and

2.1.1 RBMT

RBMT is the original approach to MT. A rule-based system assumes a large set of rules which are generally written by linguistic experts. RBMT systems are developed using one of three methodologies: direct, transfer or interlingua (Hutchins and Somers, 1992). For direct, there is very little involved in the analysis stage. A large lexicon is applied to generate a target sentence, allowing for some reorganisation but with no inherent knowledge of the syntactic relation between the source and target strings. Such approaches, therefore, are situated at the bottom of the Vauquois pyramid (cf. Figure 2.1).

In a transfer-based system, the analysis stage derives an intermediate structural representation of the source sentence. The transfer component contains a set of rules which map the intermediate representation of the source sentence onto a corresponding target representation. In the generation stage, the translation is produced from the intermediate target language representation.

The arrow in the Vauquois pyramid is positioned at an incline to indicate that in a transfer-based system the analysis stage usually requires more work than generation. For example, during analysis the intermediate representations for both the active and passive forms of a sentence may be produced. However, in generation it is possible that the MT system will only produce one form as the output string if these are deemed to be translation-equivalent.

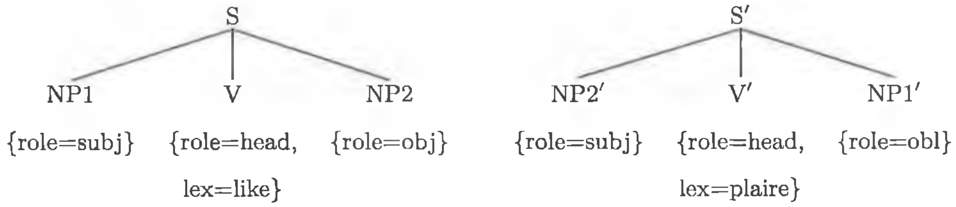
The interlingual system is located at the top of the Vauquois pyramid. This is indicative of the amount of work required by each component in a system of this type. In contrast to the transfer-based system, the stages of analysis and generation are usually more intensive. While the rules in a transfer-based system generally focus on a single language pair, the goal of an interlingual system is to abstract away from the differences in surface structure. The intermediate representations in an interlingual system, therefore, are language-independent.

In a rule-based system, transfer can sometimes be achieved by word-for-word substitution between the ⟨source, target⟩ representations. However, any major difference in structure requires the introduction of complex rules. Consider the example in (1):

(1) " Mary plaît à Jean. "
" Mary pleases to John . "
" John likes Mary."

This <English, French> example is a relation-changing case and contains a problem of 'complex transfer'. The English subject *John* acquires a dative realisation in French and the English object *Mary* becomes the subject noun phrase (NP) in French. The example in (2) is adapted from (Way, 2001) and shows possible intermediate representations for the example in (1):

(2)



In this case, performing simple transfer from source to target would result in an incorrect translation, as the object in English becomes the subject in French. In a transfer-based system, the transfer component requires detailed intricate rules to deal with such cases of complex transfer. In an interlingual system, such linguistic phenomena may be less problematic because the intermediate representations can abstract away from the level of syntax by containing semantic information. Transfer, therefore, is performed at a deeper level than surface structure.

Given that the intermediate representations in interlingual systems are language-neutral, they are more reusable than transfer-based systems, where transfer is performed on the basis of language-dependent phenomena. However, given that the analysis and generation components in an interlingual system are typically more extensive, the benefits can sometimes be outweighed by the cost of development. Furthermore, the task of creating a language-neutral representation cannot be underestimated. How does one decide what kind of representation should be encoded, given that languages vary with regard to how they deal with different linguistic phenomena?

In certain instances and in controlled environments, RBMT systems can produce good results. The METEO system (Chandioux, 1976) was designed to translate short Canadian weather reports from English into French and is an example of a successful RBMT system. However, the linguistic components required in an MT system of this nature can be expensive and cumbersome to develop. When lexical and grammatical resources need to be hand-crafted, the construction of rule-based systems can be very time-consuming. At the beginning of this chapter we noted that it is possible to automatically infer rules from corpora. This, we observed, might pave the way for rule-based techniques to overcome the 'knowledge acquisition bottleneck' and find a niche within a wider, hybrid MT paradigm. Nevertheless, it remains the case that CBMT systems boast a number of advantages over traditional rule-based approaches.

2.1.2 Advantages of CBMT

In an RBMT system, coverage of data can be difficult to achieve. Creating rules to deal with different linguistic phenomena can be complex and building on 'toy' grammars can lead to lack of robustness. The rules in an RBMT system are developed by linguists and are based on linguistic theories which are always incomplete. Therefore, they often have difficulty accepting strings which have grammatical errors or are not well-formed. Corpus-based approaches on the other hand are generally more robust than rule-based approaches and can cope with ungrammatical or ill-formed input.

In an RBMT system, sometimes adding new rules can have adverse side-effects on previously coded rules. Adding more examples to an EBMT or SMT database however, can improve the system. The potential problem of conflicting examples can be overcome by adding a weighting measure. For example, (Somers et al., 1994; Öz and Cicekli, 1998; Murata et al., 1999) apply a similarity metric which assigns a higher score to more frequently-occurring examples.

In a CBMT system, the examples are reusable. Moreover, they are real examples, providing translations in context. A corpus-based system can produce several alternative translations for a single input sentence. While some of these may be incorrect, they can be ranked and output with associated weights. For example, consider the English string *commit suicide*. An RBMT system might produce a compositional translation by translating each word individually (*commit* ⇔ *commettre*, *suicide* ⇔ *suicide*), generating the French string in (3):

(3) commit suicide ⇔ commettre suicide

A non-compositional translation would be formed if *commit suicide* was treated as a single unit of meaning associated with the French reflexive verb *se suicider.* Assuming this phrase is in the example-base, a corpus-based system could produce the correct translation *se suicider.* More importantly, even if an RBMT system does produce alternative translations, the user has no idea which one is deemed better by the system. A CBMT system will indicate its preference for one over the other. Some RBMT systems which incorporate 'preference mechanisms' (Hein, 1996) may be able to output alternative weighted translations. However, these weights are usually defined by a linguist and more often than not have no concrete empirical foundations. In contrast, the probabilities produced by an EBMT or SMT system are 'real' as they are inferred from the corpus.

An important aspect of CBMT systems is their ability to 'learn'. Regardless of the number of times an RBMT system is confronted with an input string, it will treat it in the same way each time and apply the same set of rules to derive its translation. A corpus-based system has the capacity to append new translations to its example-base. In this way, when a similar input string is encountered in a subsequent case, the translation can be retrieved directly from the example-base. CBMT systems, therefore, have the potential to 'learn' from new translations. RBMT systems have no method of integrating such information.

Another advantage of most corpus-based systems is that they do not require large-scale grammars and rules and are therefore easier to port to other languages than RBMT systems.

Finally, RBMT systems, particularly those based on the transfer approach, tend to impose the structure of the source sentence on the target translation. This is a direct contrast to the manner in which human translation is performed. Corpus-based systems are capable of avoiding such structure-preserving translation as they do not always need to encode the structural representations of the source and target sentences in the corpus.

2.2 What is CBMT?

The fundamental idea behind data-driven approaches is to generate new translations by means of a set of previously translated examples. An aligned bilingual corpus is a prerequisite. This corpus contains potentially reusable translations. CBMT systems compare new input against existing examples and extract useful fragments which are recombined to produce new translations.

Selecting a suitable corpus depends on several factors, e.g. what languages are involved, what type of corpora are available, what approach is to be pursued, etc. The quality of the corpus chosen is also an important factor, as a poor-quality training corpus is unlikely to enable the system to produce good quality translations. Another element to consider is whether the CBMT system is to be tuned to a particular sublanguage. If so, the corpus selected will need to be domain-specific. The size of the corpus also merits consideration. While it is widely acknowledged that an increased example-base can improve coverage and ultimately produce better translation results, the addition of conflicting or redundant examples can potentially have an adverse effect on translation quality (Mima et al., 1998).

2.2.1 CBMT versus Translation Memory

Another technology which has been linked with CBMT is Translation Memory (TM). While TM shares some common features with corpus-based approaches, it is a computer-assisted translation (CAT) tool and, as such, does not perform MT. Indeed, there is a fundamental difference between the two technologies. While both match new input against a set of real examples, in a TM when an exact match for an input string cannot be found in the translation database, close or fuzzy matches are retrieved and presented to the user. It is then up to the translator to locate relevant fragments and produce a target translation from these. For example, given the input in (4), a TM system might present the translator with the fuzzy matches in (5a) and (6a), so that s/he can manipulate the target strings in (5b) and (6b) to produce a final translation.

(4) While most were critical, some contributions were plain meanspirited

(5) a. While most were critical, some contributions were thoughtful and constructive

b. La plupart ont formulé des critiques, mais certains ont fait des observations réfléchies et constructives

(6) a. Others were plain meanspirited and some contained errors of fact

b. D'autres discours comportaient des propos mesquins et même des erreurs de fait

In contrast to how a TM functions, in a CBMT system the objective is to automate the matching, extraction and recombination of relevant fragments without human intervention. Furthermore, while in a TM the alignment of examples is restricted to sentence-level, CBMT systems integrate word and/or phrasal correspondences and can, therefore, facilitate the extraction and adaption of relevant fragments rather than sentences.

One aspect of TM which has contributed to their commercial success, is the amount of control which they allow the translator to retain. Any matches suggested by the TM can be freely rejected or accepted by the translator. When constructing the translations for the target document, they can use the translations proposed by the TM or alternatively ignore these translations as they see fit. They can also choose what target strings are inserted into the TM along with the source string. The translator also has the power to set a threshold for the fuzzy matching operation so that any matches below a given threshold will not be considered.

However, as illustrated in (Way and Gough, 2003), deciding on a suitable threshold can prove difficult and often involves making a sacrifice in terms of either Recall or Precision (cf. section 4.3). Consider the input in (4) and the fuzzy matches in (5a) and (6a) that might be presented to the translator. Setting the fuzzy matching threshold at 80% would mean that neither of these fuzzy matches would be retrieved from the TM as only 7/9 (77%) of the words in (5a) match those in (4) exactly, while only 3/9 (33%) of those in (6a) match those in (4) exactly. There is a danger, therefore, that setting a fuzzy matching threshold too high could result in the elimination of potentially useful matches (low Recall/high Precision). Likewise, if the threshold is set too low, it is possible that useful information could be hidden in the midst of noisy data (high Recall/low Precision).

The TM, therefore, while a popular and undoubtedly useful translation tool, stops short of MT. Although obviously comparable to corpus-based approaches, it does not extract or make use of sub-sententially-aligned fragments and does not perform any automatic recombination.

2.2.2 SMT vs EBMT

All corpus-based systems require a set of aligned ⟨source, target⟩ examples. In an SMT system (Yamada and Knight, 2001; Charniak et al., 2003), a *language model P(e)* is produced from the analysis of monolingual corpora. If the target language is English, then *P(e)* is the probability of an English string occurring. Typically, bigram and trigram models are formed. That is, only the preceding one or two words are taken into account when computing such probabilities. For example, words such as *are* and *is* are likely to occur frequently in a corpus. However, a bigram *he is* is likely to occur much more frequently than *he are.*

A separate *translation model P(e | f)* assigns the probability of *f* occurring given *e*, i.e. the probability that an English string *(e)* translates as a French string *(f)*. In an example such as that in (7), this is a fairly trivial task as each source word can be mapped sequentially to each target word.

(7) “ The girl eats. ”
“ La fille mange. ”

In a case such as that in (8), the task is more complex:

(8) “ The girl NOT eats NOT. ”
“ La fille ne mange pas. ”
“ The girl does not eat.”

In this case, *girl* maps onto fille, *the* maps onto *la* and *eats* maps onto *mange.* However, *not* in the source corresponds to two non-contiguous words in the target, *ne* and *pas*, while *does* maps onto nothing in the target. The number of words which a source word is aligned with in the target refers to the *fertility* of the word alignment. This is also calculated by the translation model, along with *distortion* (the probability of a word occurring in a particular position within a sentence).

In RBMT, a translation grammar is used to produce the target language translation. In an SMT system, given a French sentence *(f)*, the English sentence *(e)* that maximises *P(e | f)* is sought. This involves calculating the most probable result by maximising the statistical probabilities in the language model and the translation model, i.e. the probability of e occurring and the probability that f translates as e. Using Bayes rule we can rewrite the expression for the most likely translation as:

(9) $$argmax_e p(e \mid f) = argmax_e p(f \mid e)p(e)$$

Both SMT and EBMT systems integrate word-level alignments. Traditionally, while EBMT systems also made use of phrasal alignments, SMT systems were based on the original IBM approach (Brown et al., 1988) which was modelled purely on word correspondences. More recently, however, work on phrasal alignment in SMT (Och et al., 1999; Yamada and Knight, 2001; Marcu and Wong, 2002; Koehn et al., 2003; Charniak et al., 2003), has shown that including phrasal units in an SMT system can improve translation performance. As such, phrase-based SMT has become the 'norm', rather than the pure word-based systems of the original IBM model.

There are various methods for learning phrase translations in an SMT system. (Marcu and Wong, 2002) propose a method which learns phrase alignments directly from a parallel corpus. These correspondences are learned from a phrase-based joint probability model. Firstly, the joint probability that two phrases $\bar{e}$ and $\bar{f}$ are translation equivalents $\phi(\bar{e},\bar{f})$ is calculated. Secondly, a joint distribution $d(i,j)$ is determined which assigns a probability that a phrase at position i translates as a phrase at position j.

(Koehn et al., 2003) evaluate different models for phrase-based SMT. They find that phrase-based models outperform systems which use only word correspondences. As a result of their findings, a model for phrase-based translation which integrates word-based alignments and a lexical weighting of phrase translations is proposed.

(Wang, 1998; Och et al., 1999; Yamada and Knight, 2001; Aue et al., 2004) incorporate structural information into a statistical system. (Yamada and Knight, 2001) propose using a syntactic parser to process the input sentence. Word reordering is based on the order of constituents in well-formed syntactic parse trees. (Charniak et al., 2003) combine the syntax-based translation model of (Yamada and Knight, 2001) with a language model.

They report their results to be promising and conclude that both phrase-based translation and the integration of syntax improves the performance of the SMT system. (Aue et al., 2004) propose a series of models which perform SMT using labelled semantic dependency graphs. When an existing example-based system is augmented with these models, an improvement in translation quality is reported.

There are a number of reasons why EBMT systems are preferred to purely statistical ones. SMT systems require a very large bilingual corpus from which to derive probabilities. In the original IBM model (Brown et al., 1988), 30 million words of text were extracted from the ⟨English, French⟩ bilingual Hansard corpus. Most statistical systems are built on the premise that larger training corpora will lead to better results. However, when data is sparse, the statistical models may be unreliable. Furthermore, in order to assist the functionality of the system and reduce the amount of required parameter space, it is often the case that bigram models are used instead of trigram models. This, in turn, has an adverse effect on the quality of the statistical models. The reliability of an EBMT system does not depend on the probability models derived from large corpora. While some approaches to EBMT may integrate statistical techniques to order translations (Öz and Cicekli, 1998; Cicekli and Güvenir, 2003), in EBMT, the system is not entirely reliant on the probability models calculated from the training corpus. Therefore, EBMT systems can produce translations based on much smaller corpus resources.

For certain language pairs or specific domains, it may be difficult to obtain the large quantities of data required for training in an SMT system. Again, given that an EBMT system can be developed on a smaller scale this may not be so problematic if the system is example-based.

Most EBMT systems integrate at least some level of syntax. Therefore, they are more likely to produce a grammatical translation than an SMT system which contains no additional syntactic information other than what can be derived from the language model and distortion probabilities. Recent research in SMT (Yamada and Knight, 2001; Charniak et al., 2003) has focused on integrating syntactic information in SMT and there is certainly much potential for further research in this domain. However, including such information in SMT systems can require significantly more computational processing than for EBMT. Indeed, SMT systems perform complex computational processing to extract linguistic information which, in an EBMT system can often be derived more readily.

Despite the evident advantages of EBMT, in current research circles SMT seems to be more 'in vogue'. Example-based systems tend to borrow from rule-based and/or statistical techniques. Furthermore, given that SMT systems now integrate phrase-alignments, it is becoming increasingly difficult to distinguish between EBMT and SMT and most novel research in the corpus-based paradigm seems to fall into the latter category.

There seems to be a common perception that increasing the training data in a corpus-based system will improve translation performance. This may well be true for an SMT system, given that the statistical models depend on the data set. However, it is often assumed that SMT will outperform EBMT when it is trained on 'enough' data. As yet no such comparison has been made and there is no evidence to suggest that this is in fact the case.

In section 5.6, we directly compare our EBMT system with an SMT system which integrates word correspondences. As a result we show that an EBMT system can outperform an SMT system trained on similar data.

2.3 What is EBMT?

The EBMT process involves matching new data against the existing corpus, retrieving similar examples and their associated translations, and finally adapting and recombining these to produce a target translation. The diagram in Figure 2.2 illustrates the stages in EBMT.

The system is confronted with an input source sentence (Sx). The goal is to produce a target translation (Sx ⇔ Tx). In stage A (matching), the source side of the corpus is compared to the input sentence and similar examples are retrieved (S2 and S4). In stage B (retrieval/alignment), the relevant parts of the associated target fragments (T2 and T4) are identified and finally in stage C (recombination), these are recombined to produce a target translation (Tx). To illustrate this process with a real example, consider the small corpus in (10):

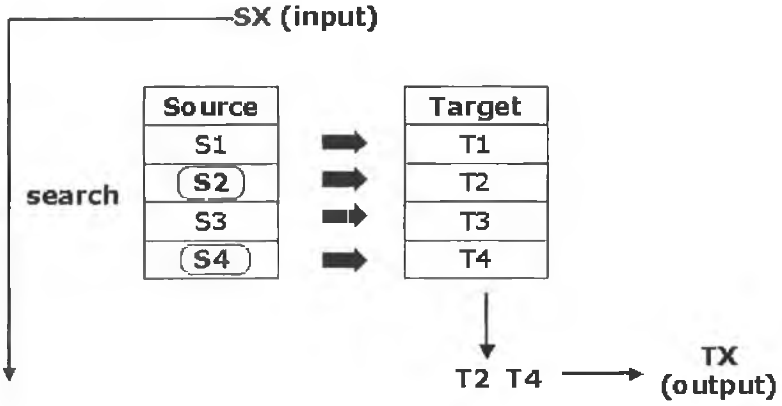


Figure 2.2: The stages of EBMT: Matching, Retrieval and Recombination.

(10) a. John went to school ⇔ Jean est allé à l'école

b. The butcher's is next to the baker's ⇔ La boucherie est à côté de la boulangerie

Assume we want to translate the English sentence *John went to the baker's* into French, via EBMT. Searching the source side of the corpus, the useful fragments in (11) may be retrieved and the corresponding target fragments are identified.

(11) a. John went to ⇔ Jean est allé à

b. the baker's ⇔ la boulangerie

Finally, the translations of the fragments are recombined to produce the French translation in (12):

(12) John went to the baker's ⇔ Jean est allé à la boulangerie

There are many dimensions to EBMT and various techniques and approaches can be applied and combined in a system which is categorised as example-based. EBMT can make use of different computational paradigms, some of which include case-based reasoning (CBR), syntax, semantics and parsing. Matching in EBMT usually involves the calculation of similarity based on words, Part-of-Speech (POS)-tags, structures, generalised templates etc. Sub-sentential alignments can be extracted from the corpus at run-time (Sumita, 2003) or in a pre-processing stage (Cicekli and Güvenir, 2003) and different systems can integrate resources such as bilingual dictionaries and thesauri (Kaji et al., 1992).

EBMT can also borrow from RBMT by applying 'rules' where it make sense to do so. The rules can be inferred automatically from the corpus. For example, the approach of (Furuse and Iida, 1992a) and (Furuse and Iida, 1992b) includes 'patterns' with substitutable variables as one of three methods of storing examples (13b). They also store literal examples (13a) and represent examples as 'context-sensitive' rewrite rules (13c).

(13) a. Sochira ni okeru ⇔ We will send it to you.
Sochira qa jimukyoku desu ⇔ This is the office

b. X o onegai shimasu ⇒ may I speak to the X (X = jimukyoku office,...)
X o onegai shimasu ⇔ please give me the X (X = bango number,....)

c. N1 N2 N3 ⇔ the N3 of the N1
(N1 = kaigi meeting, N2 = kaisai opening, N3 = kikan time)
N1 N2 N3 ⇔ N2 N3 for N1
(N1 = sanka participation, N2 = moshikomi application, N3 = yoshi form)

The approach of (Furuse and Iida, 1992a,b) is clearly corpus-based, given that it includes aligned examples. However, the similarity between the rules in (13c) and those applied in a rule-based system means that it is in fact hybrid, rather than purely example-based.

As we have already pointed out in the introduction to this chapter, the distinction between the different MT paradigms is narrowing and it is becoming increasingly difficult to distinguish between different approaches or to definitively categorise them. Attempts have been made to do so however. The work of (Turcato and Popowich, 2003) for example, analyses different approaches to MT in an effort to find common elements. The authors argue that the knowledge used is central to classifying an MT approach. They believe that the same knowledge applied in the same way by two different systems renders them equal, regardless of the approach used. Their work is an interesting and insightful attempt to define EBMT, and to narrow down its heterogenous boundaries. However, there remains no standard means of classifying an EBMT system, and approaches with varying and contrastive features are often collectively termed example-based. While it may be difficult to analytically define EBMT however, this does not imply that it cannot be an authentic discipline in its own right. In the following section we describe various approaches to

EBMT and the numerous techniques which are applied in different systems.

2.4 Related Research

In this section we describe several approaches, all of which fall within the EBMT paradigm. Many approaches integrate various linguistic resources and apply different techniques and, as such, can be difficult to categorise. Nevertheless, it is possible to associate some approaches by means of the matching algorithm applied or the manner in which sub-sentential alignments are derived. We describe related work in EBMT under the following headings:

- Word-Based Matching;
- Dynamic Programming Techniques;
- CBR techniques;
- Generalised Templates;
- Structural Approaches.

That is not to say that the approaches described in different sections cannot be linked. Moreover, some also exhibit features of statistical and rule-based approaches.

2.4.1 Word-Based Matching

(Nagao, 1984) proposed using thesauri to indicate word similarity on the basis of meaning or usage. A thesaurus provides a listing of synonyms, allowing examples to match the input, on condition that they can be classified as synonyms based on a measurement of similarity. The examples in (14) and the translations in (15) from (Nagao, 1984) show how this technique can be used successfully in choosing between conflicting examples.

(14) a. A man eats vegetables ⇔ Hito wa yasai o taberu

b. Acid eats metal ⇔ San wa kinzoku o okasu

(15) a. He eats potatoes ⇔ Kare wa jagaimo o taberu

b. Sulphuric acid eats iron ⇔ Ryusan wa tetsu o okasu

In (15a), the correct translation of *eats (taberu)* is chosen. This is correct in this instance as it refers to food and is chosen because of the relative similarity or distance between *potatoes* and *vegetables*. Similarly *okasu* is correctly selected as the translation of *eats* in the context of (15b).

The above example illustrates how this technique can be used to resolve lexical transfer ambiguity. It has also been applied in other EBMT systems to overcome structural transfer ambiguity. One example is that of (Sumita et al., 1990) and (Sumita and Iida, 1991) where EBMT is applied effectively for translating Japanese adnominal particle constructions *(A no B)*. When translating such constructions, a large number of translation patterns can potentially be generated and producing the structure-preserving translation *(B of A)* proves to be incorrect 80% of the time. However, by using a thesaurus to measure the similarity between words the correct translations can be produced.

Given the example pairs in (16) and the Japanese sentences in (17) from (Somers, 2003), a partial match can be established and a thesaurus can relate *Kyoto* and *Tokyo* as they are both place names. In a similar manner, *kaigi* (conference) and *kenkyukai* (workshop), and *densha* (train) and *shinkansen* (bullet train) can be related and the English translations in (17) can be produced by substituting semantically similar words in the partial match.

(16) a. *kyōto-de no kaigi*
KYOTO-IN adn CONFERENCE
a conference in Kyoto

b. *kyōto-e no densha*
KYOTO-TO adn TRAIN
the Kyoto train

(17) a. *tōkyō-de no kenkyukai*
a workshop in Tokyo

b. *tōkyō-e no shinkansen*
the Tokyo bullet-train

The approach to matching described above is based on finding similar words in existing examples. Other techniques involve matching based on annotated words. For example, (Cranias et al., 1994, 1997) match against existing examples using POS-tags and function words. The marker-based approach of (Veale and Way, 1997) also matches input sentences based on similar function words. This approach is described in more detail in section (2.5). In our approach, we use function words to segment the examples in our training data and derive a set of additional sub-sententially-aligned lexical resources. However, unlike (Veale and Way, 1997), our matching algorithm is based on the location of *n*-grams within our training corpus.

2.4.2 Applying Dynamic Programming Techniques

(Sumita, 2003) applies an algorithm based on dynamic programming (DP)-matching between word sequences for a speech-to-speech translation system. DP techniques provide optimal solutions to specific problems by making decisions at discrete time stages. At each stage, a small number of finite options are possible. Decisions are made based on obtaining the optimal path from the input sentence to an example sentence.

(Sato, 1993) and (Cranias et al., 1997) have also applied dynamic programming techniques in EBMT. (Sato, 1993) translates technical terms and derives a matching score using DP techniques. (Cranias et al., 1997) apply DP to a TM in an effort to improve retrieval using clustering techniques.

In Sumita's approach, retrieval of examples is based on the calculation of a distance measure *(dist)* between the input and the example sentences. This distance measure is a normalised score of the sum of substitutions, deletions (D) and insertion (I) operations and is calculated using the formula in (18). *SEMDIST* refers to the semantic distance between two substituted words and is calculated using a thesaurus.

(18) $dist = \frac{I+D+2\sum SEMDIST}{Linput+Lexample}$

Once a similar example has been detected, the next step is to formulate a translation pattern from this example. These patterns are created dynamically and are not retained or stored for use in future translation. To illustrate how this works, assume the Japanese input sentence in (19):

(19) " *iro* ga ki ni iri masen. "
" *colour* SUB favour OBJ enter POLITE-NOT. "
" I do not care for the colour."

Given the sentence in (19), the example in (20) might be retrieved, based on its similarity to the input.

(20) " *dezain* ga ki ni iri masen. "
" *design* SUB favour OBJ enter POLITE-NOT. "
" I do not care for the design"

In the retrieved translation example, the italicised words, *dezain* and *design* can be aligned and the translation pattern in (21) is obtained by replacing differing portions with variables. The differing portions (Xj and Xe) can then be aligned. The remainder (underlined) is treated as an indivisible unit and is not aligned word-for-word.

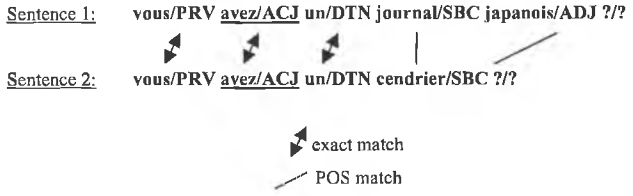
(21) Xj/ga/ki/ni/iri/masen ⇔ I do not like the Xe

In the translation of new input, a bilingual dictionary is exploited to replace target variables with words in the target language. In the above example, this would be achieved by using the target part of the translation pattern in (21) (*I do not like the Xe*) and replacing the variable *Xe* with the English translation of *iro*. This produces the final translation in (22):

(22) I do not like the colour.

In (Andriamanankasina et al., 2003), the matching algorithm is partially similar to the DP-matching algorithm presented by Sumita. However, while the DP-matching algorithm starts a search at one end of a segment and continues until the start of another, the matching algorithm presented here involves taking each of the words in the input example and searching for examples that contain each of these words. Where a word is located, a search for matches before and after this word ensues. A similarity score is applied to locate the best-matching sentences. This score is based on the number of exact matches and the number of matching POS-tags. The distance between the original segment located and an exact match is also taken into account, with those found closer to the original common segment gaining higher weighting.

Figure 2.3 illustrates the matching process. The initial match is located at *avez/ACJ* at the second position in both sentences. A backward search matches *vous/PRV* ⇔ *vous/PRV* and a forwards search matches *un/DTN* ⇔ *un/DTN*. As *journal/SBC* and *cendrier/SBC* are not an exact match, only the POS-tags are matched hereinafter.


Sentence 1:
vous/PRV avez/ACJ un/DTN journal/SBC japanois/ADJ ?/?
Sentence 2:
vous/PRV avez/ACJ un/DTN cendrier/SBC ?/?
exact match
POS match

" vous avez un journal/cendrier japonois . "
" you have a newspaper/ashtray japanese . "

" do you have a japanese newspaper/ashtray ."

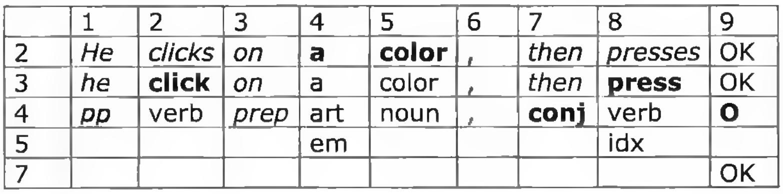
Figure 2.3: Matching in the translation method of (Andriamanankasina et al., 2003)

A POS-tagged, word-aligned corpus is presupposed. The system has the ability to learn, in that new examples can be integrated into its knowledge base. Moreover, links between words in new examples can be established by a link prediction module. Aligning the corpus initially can be cumbersome, particularly if manual alignment is required. The link prediction module is intended to automatically specify correspondences between words in a new example pair and is implemented via a CBR (cf. section 2.4.3) technique. Pairs of existing manually-linked examples are chosen and links between new source and target elements are learned from these.

(Planas and Furuse, 2003) store examples in a multi-layer structure (TELA). Each layer of the TELA structure contains different forms, i.e. words, lemmas, POS-tags etc. The original sentences are extracted from a TM. For example, (24) shows the encoding of the sentence in (23) in the TM. The corresponding simplified TELA structure is shown in Figure 2.4.

(23) He clicks on a color, then presses OK.

(24) <s>He_clicks_on_a_<em>color,_</em>then_
<idx attrib=01> presses </idx>_OK</s>



	1	2	3	4	5	6	7	8	9
2	*He*	*clicks*	*on*	**a**	**color**	*,*	*then*	*presses*	OK
3	*he*	**click**	*on*	a	color	*,*	*then*	**press**	OK
4	*pp*	verb	*prep*	art	noun	*,*	**conj**	verb	**O**
5				em				idx	
7									OK

Figure 2.4: A simplified TELA structure for the sentence in (23)

A TELA structure can have as many layers as necessary. The bottom row of each layer in the TELA structure is less important than all other elements of that layer. Similarly, the top row is the most important row. The first row (not shown in the simplified Figure 2.4) contains relevant characters in the text. The second row contains the surface form of the words in the sentence. The third and fourth rows contain lemmas and POS and the fifth row shows Extensible Markup Language (XML) content tags present in the TM. Row six (not shown in the simplified Figure 2.4) contains any empty tags to cope with objects such as images inserted in the XML text. Finally, row seven contains any glossary entries which specify terminology information for the current context.

Matching an input TELA structure to the TELA structures of the existing examples is achieved by locating similar sentences using an index-based technique and subsequently applying a Multi-level Similar Segment Matching (MSSM) Algorithm based on dynamic programming techniques, to refine the original set of examples retrieved. The original algorithm developed by (Wagner and Fischer, 1974) finds a match from a node C to a node I based on edit distance and accounts for deletions, equalities and substitutions between examples. (Planas and Furuse, 2003) use an adapted version of this algorithm to only account for deletions and equalities between examples and the input, therefore eliminating the retrieval of fuzzy matches. The number of equalities and deletions is calculated at each layer of the TELA structure.

A bilingual dictionary is used to establish initial lexical correspondences. The nature of the dynamic programming MSSM algorithm allows it to back-track so that the position of corresponding words can be found. In this way a 'trace' of word correspondences is also produced, giving rise to lexical equivalences derived from ⟨source, target⟩ examples.

2.4.3 Applying CBR techniques

CBR applies past cases to solve new problems. Typically, each case contains a description of the problem and a possible solution. The case-based *ReVerb* system (Collins, 1998) applies CBR techniques to EBMT. In this approach, candidate examples are initially selected on condition that they share n words with the input. From this set, a parsed representation of each example is compared against a parsed representation of the input. This is an attempt to locate a match based on syntactic function. Failing this, syntactic function is combined with the additional parameters of sentence position and lexical equivalences. Where more than one match has been retrieved at this stage, matches are scored in terms of *adaptability*. This is a term borrowed from the CBR paradigm and refers to the extent to which an example needs to be adapted to form the desired output. A balance between similarity to the input and adaptability culminates in the retrieval of a single example.

Word equivalences are determined by exploiting the corpus. A bilingual dictionary is constructed from the example-base on the assumption that co-occurring source and target words are likely to be translations of each other.

Consider the example from (Collins, 1998). The input sentence in (25) shares similarities with both the English sentence in (26) and the English sentence in (27). However, as illustrated in Figure 2.5, the German sentence in (27) is more adaptable in this case. The ⟨source, target⟩ fragments in (27) are more explicitly linked and it is more obvious what the structure and word order of the target sentence should be.

(25) Use the Offset Command to increase the spacing between the shapes

(26) " Mit der Option Abstand legen Sie den Abstand zwischen den Formen
" with the offset command make you the spacing between the shapes
fest. "
firm. "

" Use the Offset Command to specify the spacing between the shapes. "

(27) "Mit der Option Speichern können Sie ihre Anderungen auf Diskette
"with the save option can you your changes to disk
speichern . "
save . "

"Use the Save Option to save your changes to disk. "

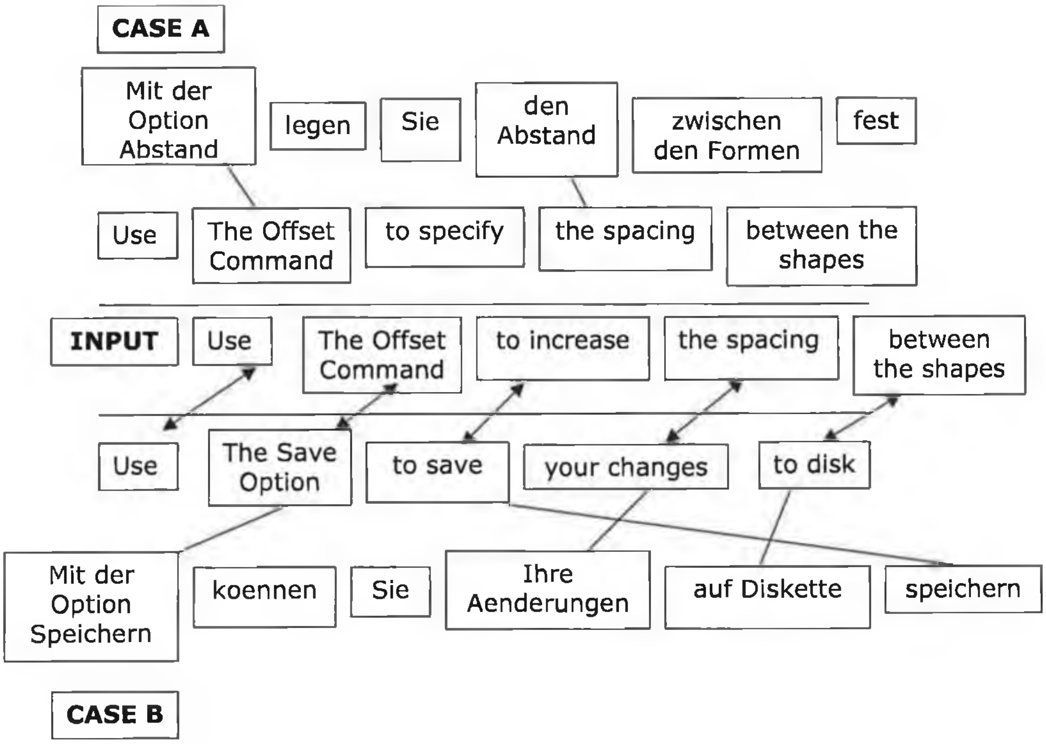


Figure 2.5: Adaptability versus similarity in retrieval

In their hybrid approach, (Bond and Shirai, 2003) also apply CBR techniques in their matching algorithm. Firstly, examples are retrieved based on n-gram matching techniques. A similarity metric based on the order of shared segments and their co-occurrence ranks these and clustering techniques choose the most adaptable example. Those sections of the retrieved example which differ from the input sentence are identified. The hybrid nature of this approach allows rule-based techniques to translate these differing portions and insert the translations produced into the target translation.

2.4.4 Generalised Templates

While (Sumita, 2003) creates translation patterns or templates dynamically at run time, other systems extract these templates from the example-base prior to the translation process. This typically involves generalising the existing example pairs by replacing similar and/or differing portions with a variable. The variables contained in these generalised templates are then instantiated with fragments corresponding to the target translation. The extent to which generalisation is carried out differs from system to system.

In their approach, (Kaji et al., 1992) employ significant linguistic resources and generalise by syntactic category. They use Japanese and English parsers to extract syntactic categories and subsequently align these using a bilingual dictionary. Any aligned pair can then be replaced by a variable. In this way, the templatised example in (29), can be produced from the example pair in (28), as *Rekodo* is aligned with *record* and *512* is aligned with *512*. Of course, any other words in the source and target sentence which are aligned using the bilingual dictionary can also be replaced by a variable.

(28) Rekodo no nagase wa saidai 512 baito de aru ⇔
The maximum length of a record is 512 bytes

(29) X[NP] no nagasa wa saidai Y[N] baito de aru ⇔
The maximum length of a X[NP] is Y[N] bytes

Where templates conflict, semantic categories are used to refine them. For example, in Japanese, the verb *play* can have different translations, depending on the context in which it appears. The examples in (30) illustrate this.

(30) play the piano ⇔ piano o hiku
play tennis ⇔ tenisu o suru

Adding semantic categories to the translation templates as in (31) can eliminate this ambiguity by specifying the context in which each form of the verb can occur.

(31) play X[NP/sport] ⇔ X[NP] o suru
play X[NP/instrument] ⇔ X[NP] o hiku

(Cicekli and Güvenir, 2003) present an approach (English-Turkish) which uses a morphologically tagged aligned bilingual corpus. While their earlier work attempted to construct parse trees, such an approach was abandoned when difficulties were encountered in locating reliable parsers. In order to overcome the limitations of an agglutinative language such as Turkish in terms of generality, words are represented at a lexical level (i.e. as stems and morphemes). Two heuristics, based on analogical reasoning and mirroring human language acquisition, are used to identify similar and different portions between existing examples. These are implemented as the TTL (Translation Template Learner) algorithms. When similar and differing portions are identified, templates are created by replacing these portions with variables. The templates are then sorted in accordance to how specific they are. Previously learned templates can be applied to help learn templates from new examples. The examples in (32) illustrate the generalisation process, given the translation pairs in English and Turkish:

(32) I will drink orange juice ⇔ portakal suyu içeceğim
I will drink coffee ⇔ kahve içeceğim

The similarities between the examples are underlined and the remaining parts are the differences. Replacing the differing portions with relevant variables gives the templates in (33):

(33) I will drink Xe ⇔ Xt içeceğim

Substituted variables in translation templates are linked to establish correspondences between them. As well as storing the generalised templates formed, atomic templates — representing aligned strings in the examples and devoid of variables — are also retained. Therefore, from the example pair in (32), the sub-sentential alignments in (34) can also be derived:

(34) a. orange juice ⇔ portakal suyu

b. coffee ⇔ kahve

(Brown, 1999) creates templates by replacing words with tokens, which are an indication of what class of word can occur in that position. Assume the English sentence in (35):

(35) John Hancock was in Philadelphia on July 4th

Brown's approach generalises the sentence by replacing the words representing a person, a city and a date by a label indicating their word category. These words are known as 'placeables'. This results in the generalised template in (36):

(36) <PERSON> was in <CITY> on <DATE>

In more recent work, (Brown, 2003) also adopts a template-driven approach based on 'purist' EBMT. Aside from the bilingual dictionary, all lexical information is extracted from the parallel aligned corpus. Brown combines the induction of transfer rules, as demonstrated by (Cicekli and Güvenir, 2003), with word-clustering techniques of previous work (Brown, 2002) to produce equivalence classes of single words and transfer rules. These can be stored for use in translating new sentences.

Cluster	***French***	***English***
507	NE NE	NOT NO
568	CETTE CETTE	THIS THAT
2609	<CL_{54}> <CL_{98}> <CL_{375}> <CL_{458}> <CL_{462}>	<CL_{54}> <CL_{98}> <CL_{375}> <CL_{458}> <CL_{462}>
1776	ABSURDE <CL_{18}>	NONSENSE <CL_{18}>

Table 2.1: Sample Clusters from 107,000 words in the approach of (Brown, 2003)

For instance, consider the following example adapted from (Brown, 2003). English and French words such as those in Table 2.1 can be clustered.

Equivalence classes such as $<CL_{54}>$ and $<CL_{98}>$ can be applied in similar contexts and therefore it is also useful to cluster these. Brown applies the same word-clustering techniques to cluster equivalence classes. In a final stage, words and equivalence classes can be clustered together.

In our approach, we cluster on certain function words, thus increasing the flexibility of our matching algorithm and ultimately the coverage of our system (cf. 3.2.4).

(McTait, 2003) forms translation patterns which to some extent mimic the transfer rules found in rule-based systems. However, they are less restricted. This approach is based on analogical reasoning, and as in the case of both (Brown, 2003) and (Cicekli and Güvenir, 2003), McTait supports a language-independent methodology. In this respect, he chooses not to avail of cognate matching or other language-dependent phenomena. Translation patterns are extracted in a similar manner to the approaches described above.

When establishing correspondences between fragments in source and target sentences, it is not always the case that each word in the source can be mapped 1:1 to each word in the target. (McTait, 2003) allows variables to have 1:2 mappings so that a word sequence in an English sentence such as *gave up* can be correctly mapped to its French translation *abandonna*. The algorithm is based on co-occurrence and applies a frequency threshold. That is to say, if source language or target language strings co-occur in two or more examples, they can be deemed translations if the distance between them does not rise above a given threshold.

Collocations of co-occurring lexical items can be formed through a recursive combination of individual lexical items. This process accounts for the monolingual phase of extracting translation patterns. For example, given the corresponding ⟨English, French⟩ sentences in (37), the information in (38) can be extracted:

(37) a. The commission **gave** the plan **up** ⇔ La commission **abandonna** le plan

b. Our government **gave** all laws **up** ⇔ Notre gouvernement **abandonna** toutes les lois

(38) a. (gave)[37a,37b](up)[37a,37b]

b. (abandonna)[37a,37b]

If a lexical item occurs in two or more sentences, then it is retrieved along with the index (37a,37b) of each sentence in which it is found. In (38) the lexical items *gave* and *up* are allowed to combine as they share two common sentence indexes (37a and 37b). By assuming that the collocations formed in target language examples provide the translations of similarly indexed collocations in source language examples, translation patterns such as that in (39) can be generated. (The bracketed ellipses represent variables).

(39) (...) gave (...) up ↔ (...) abandonna (...)

McTait adds varying levels of linguistic information to attempt to improve the performance of the system but suggests that such additional information may reduce the portability of the system to other language pairs and therefore may ultimately have an adverse effect.

(Block, 2000) bases his generalisation techniques on the assumption that the only available knowledge source is the aligned corpus. He does not presuppose the existence of other linguistic resources. In Block's approach, word alignments are generated using statistical techniques. This information is used as a basis for extracting 'chunk pairs'. Generalisations are created by replacing selected chunk pairs with variables. For example, given the word-aligned ⟨German-English⟩ sentence pair in (40), the chunk pairs in (41) can be produced among others.

(40) das\is\was\Sie\sagten
which\is\what\you\said

(41) a. das ⇔ which

b. ist ⇔ is

c. das ist ⇔ which is

d. ist was ⇔ is what

e. das ist was ⇔ which is what

f. ist was Sie ⇔ is what you

g. ....

Generalisations (or as termed by Block, 'pattern pairs') are formed by replacing the chunk in the first string with a variable *V*, and assigning its translation the same variable. For example, according to Block's methodology, the string *das* in (41a) is a substring of *das ist* in (41c) , and therefore can be replaced by a variable as in (42):

(42) a. V ist ⇔ V is

b. V was ⇔ V what

c. V ist was ⇔ V is what

d. V was Sie ⇔ V what you

e. das V ⇔ which V

f. ist V ⇔ is V

g. das V was ⇔ which V what

h. ist V Sie ⇔ is V you

i. ....

The approach of (Carl, 2003b) shares some similarities with the work of (Block, 2000). However, while Block uses a statistical word aligner, this approach uses a shallow parser to induce bracketed alignments. Also, Block is limited to replacing only one pattern pair in a generalisation. (Carl, 2003b) requires a dictionary along with morphologically tagged, lemmatized and bracketed alignments. To create templates, the alignments are substituted with variables and morphological information is attached. For example, given the (German,English) aligned pair in (43), the templates in (44) can be generated.

(43) Hans sieht den Mann mit dem Fernglas ⇔ John sees the man with the binoculars

(44) $(\{noun\}^1$ sieht $\{dp\}^2) \leftrightarrow (\{noun\}^1$ sees $\{dp\}^2)$
$(\{noun\}^1$ sieht $\{dp\}^2$ $\{pp\}^3) \leftrightarrow (\{noun\}^1$ sees $\{dp\}^2$ $\{pp\}^3)$

The alignment in (43) is ambiguous. The sentence *John sees the man with the binoculars* can be interpreted as:

- John sees the man who has the binoculars;
- John sees the man by looking through his binoculars.

Despite the ambiguity present in the sentence, the alignment is correctly bracketed and therefore two possible templates are produced.

In our marker-based approach, we also derive generalised templates from our training corpus. We use a technique similar to that of (Block, 2000). However, in contrast to Blocks' method, we do not apply a statistical word alignment tool, but instead replace certain function words with variables (cf. section 3.2.4).

2.4.5 Structural Approaches

Traditional approaches to EBMT represent examples as tree structures with linguistic information attached (Sato and Nagao, 1990; Watanabe, 1992; Sato, 1995). Parsed input is matched against the structurally composed examples. A target language tree is formed from relevant fragments and a translation is generated from this.

Statistical parsing techniques can derive dependency structures. These can provide the basis for forming word and phrasal correspondences between source and target examples. In the approach of (Yamamoto and Matsumoto, 2003), the initial experiment reports 90% accuracy when using a dependency parser for this purpose. The authors then combine NLP techniques with linguistic knowledge derived from dependency structures, in the hope that this will reap benefits for the extraction of translation units.

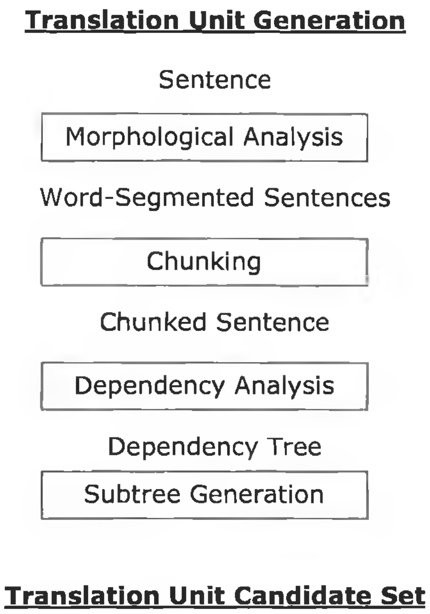


Figure 2.6: Generation of corresponding Translation Units (Yamamoto and Matsumoto, 2003)

Figure 2.6 illustrates the various experiments which are carried out.. Three separate models are used for the purpose of investigating how using different linguistic clues is linked with the quality of the translation knowledge extracted. These include bounded-length n-grams, chunk-bounded n-grams and dependency-linked n-grams.

(Yamamoto and Matsumoto, 2003) report that chunk-bounded and word dependency-linked n-grams produce better results than the baseline plain n-gram model. They conclude therefore, that word dependencies and chunk boundaries are useful for extracting translation knowledge. The experiment using chunk-bounded n-grams produced the best results, due to the increased reliability of the shallow parser in comparison with the dependency parser. However, the authors also emphasise that links produced from the dependency parser can be used successfully in the translation of domain-specific or idiomatic expressions.

In the experiment of (Yamamoto and Matsumoto, 2003), no bilingual dictionary is used and there is no distinction made between word and phrasal correspondences. In contrast, (Watanabe et al., 2003) apply a bilingual dictionary and distance measure in their approach, to provide an initial set of word correspondences. Phrasal correspondences are derived from here. The authors suggest that parsing errors could be overcome by some

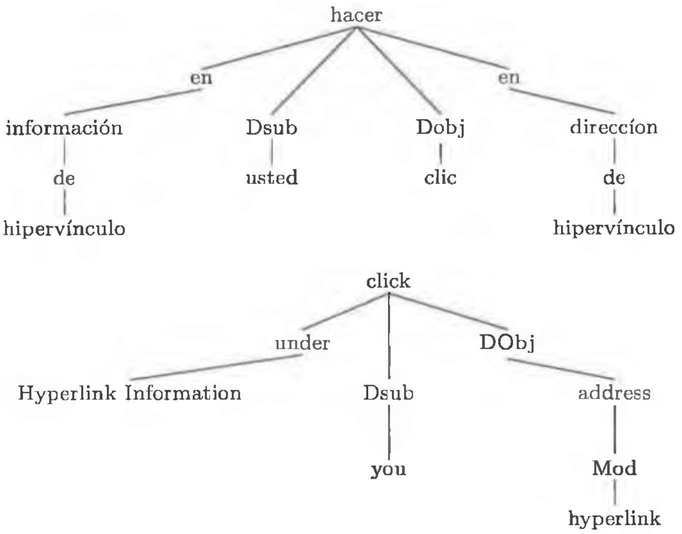


Figure 2.7: Spanish and English LFs for example sentences in (45) under the approach of (Menezes and Richardson, 2003)

manual correction.

(Menezes and Richardson, 2003) also make use of a bilingual dictionary along with statistical techniques to extract correspondences between words in source and target sentences. Rule-based parsing techniques construct dependency structures from the sentences in an aligned bilingual corpus. These structures are referred to as Logical Forms (LFs), which abstract away from language-specific phenomena. For example, from the <English,Spanish> sentence pair in (45), we can acquire the LFs in Figure 2.7.

(45) En Información del hipervínculo, haga clic en la dirección del hipervínculo ⇔
Under Hyperlink Information, click the hyperlink address

Nodes in the LF structures represent words. Alignment candidates or lexical correspondences are initially selected by means of a bilingual dictionary and statistical techniques. Using a best-first strategy implemented via a translation grammar, nodes are aligned. In this process, the best or most unambiguous lexical correspondence is located. Working outwards from that point, alignments can be determined. These are subsequently upgraded to transfer mappings by appending context, and frequencies are attached to these. For example, from the LFs in Figure 2.7 the transfer mappings in Figure 2.8 are acquired.

The work presented by (Poutsma, 2003; Way, 2001, 2003) uses a similar approach to those described above. (Way, 2003) illustrates hybrid EBMT using different combinations

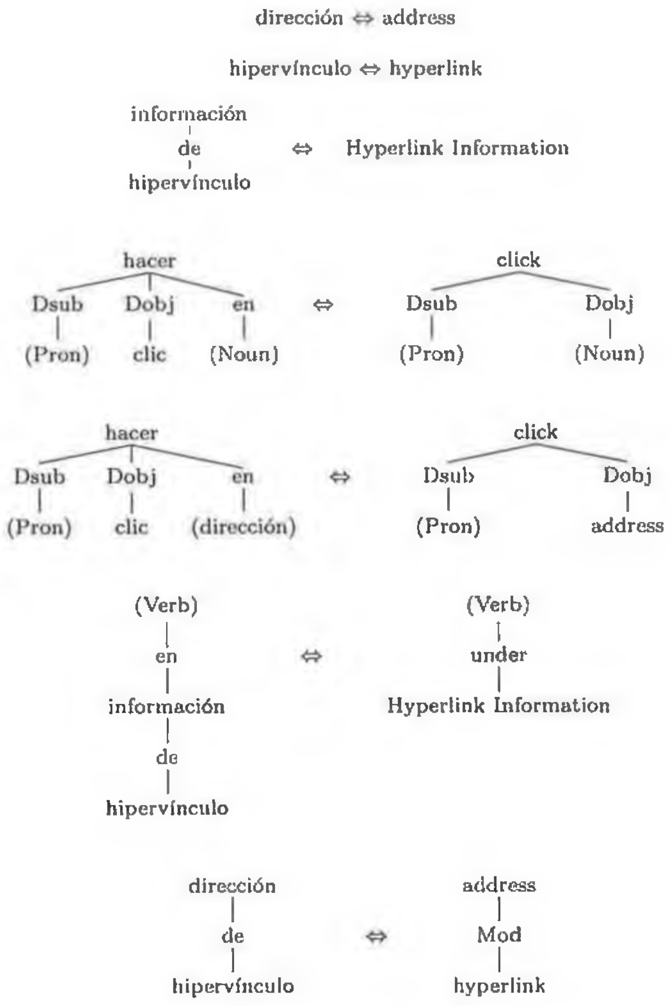


Figure 2.8: Transfer mappings acquired from Spanish and English LFs in (Menezes and Richardson, 2003)

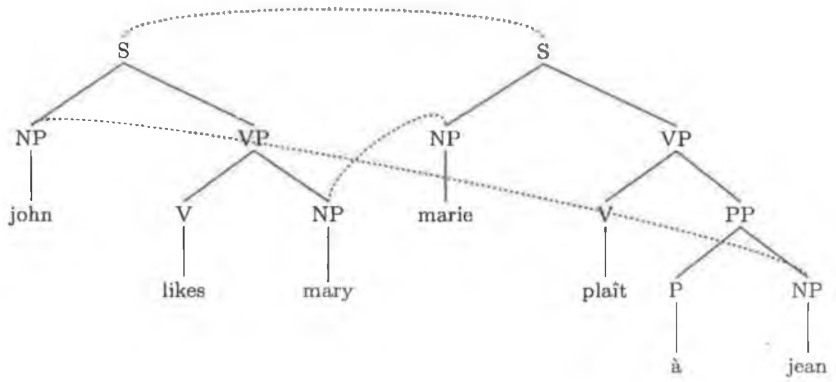


Figure 2.9: 'Linked Phrase-Structure Trees in DOT'

of Data-oriented parsing (DOP) and Lexical Functional Grammar (LFG). Altogether, four models are discussed. The work of (Poutsma, 2003; Hearne and Way, 2003) implements a similar model but structures are devoid of LFG annotations. As shown in Figure 2.9, examples are represented as annotated phrase structure trees. As in previous approaches there are explicit links between source and target correspondences. The DOT model of (Hearne and Way, 2003) is data-driven and integrates a combination of statistical, linguistic and example-based techniques.

While structural approaches may be more likely to produce a grammatical sentence, they generally require extensive linguistic and computational resources. The approach which is pursued in this thesis can be contrasted with the approaches described in section 2.4.5. Rather than integrating parsers to create detailed structural representations we apply a shallow parsing technique facilitated by the Marker Hypothesis (c.f section 3.2.2).

2.4.6 Recombination and Boundary Friction

Having matched examples and retrieved relevant fragments from the example-base, the next step is to produce a translation by recombining fragments in an appropriate fashion. Ideally, this should result in the production of a grammatically valid target sentence, which corresponds to the input sentence. However, simply pasting fragments together can often produce errors in the complete translation formed, particularly in languages that are highly inflectional, and boundary friction is common. Boundary friction can occur when fragments of translations are extracted from different contexts. When these are pieced together the string produced may not be well-formed. For example, recall the input sentences in (10) *John went to the baker's.* Assume that the fragments extracted from the ⟨English, French⟩ corpus for this sentence are those in (46):

(46) a. John went to ⇔ Jean est allé à

b. the ⇔ le

c. baker's ⇔ boulangerie

When these are recombined, the string in (47) is produced:

(47) Jean est allé à le boulangerie

The masculine determiner *le* does not agree with the feminine noun *boulangerie* and therefore this example suffers from boundary friction.

Different approaches to EBMT have experimented with various techniques for overcoming such problems. Linguistically rich structural approaches where fragments are explicitly linked are less likely to suffer from problems of boundary friction. This is demonstrated in the work of (Way, 2003). Way shows that the syntactic information inherent in the f-structures, reduces the problem of boundary friction.

Purely statistical techniques invoke a 'language model' at the recombination stage to maximise the product of the word-sequence probabilities. A method of measuring the 'correctness' of proposed translations in EBMT can be derived from statistical techniques where n-gram frequencies can be determined using a large corpus. One such experiment was performed by (Grefenstette, 1999), who used the WWW to verify alternative translations of ambiguous noun compounds. One example given is of the French compound *group de travail* which literally means *group of work.* The correct English translation of this expression is *work group.* When a Web search is performed, *work group* obtains 67,328 hits, significantly higher than some of the alternative erroneous translations produced, such as *labour grouping* and *labour group* which received 4 and 844 Web hits respectively.

Rule-based techniques can also be integrated *post hoc* to check the grammaticality of the translation. In the hybrid approach of (Bond and Shirai, 2003) (cf. section 2.4.3), the resulting translation is analysed at surface level to check for boundary friction and a target language grammar is used to smooth over the resulting translation.

Another method of overcoming boundary friction is tested in (Somers et al., 1994). Here 'hooks' are used to signify left and right context. These indicate which words and POS-tags can occur in the immediate left and right context of a fragment, thus limiting the extent to which incorrect information can be inserted. (Somers et al., 1994) compare candidate target translations against the target examples in the corpus. The existing examples are real and assumed to be of good quality. Therefore, the proposed target translation exists somewhere within the example-base, so this may be further evidence that it is a well-formed string.

The approach of (Andriamanankasina et al., 2003) described in section 2.4.2 integrates a similar technique in its matching algorithm. (McTait, 2003) (cf. section 2.4.4) also checks for overlapping sequences of up to five words between the target fragments attached to each variable and the target side of the corpus. For example, given the English input sentence in (48), the template in (49) is retrieved.

(48) AIDS control programme for Ethiopia

(49) Aids control programme for (...) ⇔ program contra el SIDA para (...)

The word *Etiopía* can be inserted in the target side of the template in place of the variable (...). A search is then performed to attempt to locate the word sequence (programa, contra, el, SIDA, para) in the target side of the corpus before *Etiopía*. The more words from this sequence which can be located before *Etiopía*, the higher the score assigned to the overlap.

(Brown, 2003) demonstrate a similar approach which locates overlapping fragments. In Brown's approach, overlapping source fragments are also taken into account. When adjacent source and target fragments overlap, the combination of these fragments is more likely to be an accurate translation.

As is the case with most EBMT systems, we too suffer from problems of boundary friction. The approach which we have applied to address this issue is based on the technique implemented by (Grefenstette, 1999) and is described in detail in section 3.5.

2.5 Marker-Based EBMT

The concept of identifying 'linguistic universals' has been explored by researchers with some success (Greenberg, 1966; Berlin and Kay, 1969; Keenan and Comrie, 1977). (Chomsky, 1981) proposed the notion of a 'universal grammar' which could limit the set of languages in the world. Chomsky theorised that by describing languages in terms of the rules and constraints that can be applied to them, a small set of properties can be derived from a potentially large set of specific rules. By applying this theory to MT, one can reduce the complexity of the knowledge acquisition problem. However, without extensive linguistic analysis, it can be difficult to define and apply such properties.

The Marker Hypothesis (Green, 1979) is a psycholinguistic constraint which states that all natural languages are 'marked' for complex syntactic structure at surface form by a closed set of specific lexemes and morphemes. In our approach, the Marker Hypothesis provides the basis for the segmentation of example-pairs and the subsequent derivation of smaller aligned fragments. In this section, we will describe how the Marker Hypothesis has been applied successfully in various NLP applications and more specifically in the area of MT.

The Marker Hypothesis has been applied in numerous language-related applications, including:

- language learning (Green, 1979; Mori and Moeser, 1983; Morgan et al., 1989);
- monolingual grammar induction (Juola, 1998);
- grammar optimization (Juola, 1994);
- insights into universal grammar (Juola, 1998);
- machine translation (Juola, 1994, 1997; Veale and Way, 1997).

With regard to translation, one problem that might be envisaged when applying the Marker Hypothesis is the non-existence of marker words such as articles in some languages. For example, the English phrase *a small boy* translates as *buachaill beag* in Irish. The noun *boy* translates as *buachaill* and the adjective *small* translates as *beag*, but the translation of the article *a* is not required.

However, (Green, 1979) showed that psycholinguistic cues based on the Marker Hypothesis theory facilitated the acquisition of artificial languages both with and without specific marker words. Similarly (Mori and Moeser, 1983) showed that case marking on pseudo-words in such artificial languages aided the language learning process. (Morgan et al., 1989) demonstrated that the Marker Hypothesis can be successfully applied to languages where phrases cannot be substituted by pronouns. (Juola, 1994) points out that there is also typological evidence to support the Marker Hypothesis. Even for simple languages without grammatical affectation, such cues exist. To give an example, the pidgin language *Russenorsk* marks verbs with 'om'.

Marker words and morphemes have been shown to be universal but also to demonstrate similarity across languages. (Talmy, 1988) suggests that some concepts are expressed grammatically (via markers or structural cues) and some lexically. For instance, in many languages nouns are inflected to express number and number is therefore expressed grammatically. For example, the translation of *one girl* from English into French is *une fille.* The translation of *two girls*, however, is *deux filles.* On the other hand, the concept of colour is not expressed grammatically, as no morpheme is used in any language to differentiate between nouns of different colours (red dress, blue dress, pink dress etc.,). According to (Juola, 1994), this is evidence that marker constructs do exist universally across all languages and that the semantic concepts that they represent can be expressed in different languages by different marker constructions.

(Juola, 1994, 1998) applies the concept of the Marker Hypothesis to grammar induction and optimisation, showing how context-free grammars (CFGs) can be converted to *marker-normal form.* This psycholinguistic model states that no constituent can be unmarked, or that "no CFG production has two adjacent nonterminal symbols on the right-hand side". For a CFG grammar to be in *marker-normal form*, all of its rules must be one of the types listed in (50), where a non-terminal (a marker word) is represented by (A) and a terminal symbol (a non-marker word) is represented by (a).

(50) $A \rightarrow \epsilon$,

$A \rightarrow a$,

$A \rightarrow A_0\ a_1\ A_1\ a_2\ A_2$ .....

$A \rightarrow a_1\ A_1\ a_2\ A_2$ .....

For languages which do not have a one-to-one mapping between a terminal symbol and a word, marker-normal form grammars fail to capture regularities. However (Juola, 1998) deals with this problem by allowing for a 'slightly more general mapping'. Adjacent terminal symbols (such as a word and its case-marking) can be merged into a single lexical item.

While these mappings are of a monolingual nature, they are extended to language pairs in Juola's work on MT (Juola, 1994, 1997). Juola's METLA system is based on the Marker Hypothesis. It uses an aligned bilingual corpus to 'infer' a source language grammar. The source sentences are then parsed and rewritten in the target language.

Some small experiments were conducted from English-French and English-Urdu using METLA. For the former language pair, 61% accuracy is reported when the system is tested on the training corpus, and 36% when evaluated with test data. For English-Urdu, (Juola, 1997) reports 100% accuracy when tested on the training corpus and 72% accuracy when tested using novel sentences.

In their *Gaijin* system, Veale and Way (1997) also apply the Marker Hypothesis to produce sub-sentential alignments and templates with replaceable variables. When translating from English-German, they report 63% accuracy on a testset of 791 sentences derived from *Corel Draw* manuals.

In the *Gaijin* system, a bilingual sentence alignment algorithm is initially applied to produce aligned ⟨source, target⟩ examples from a bilingual corpus. The source and target words are then related to each other via a correspondence matrix. This is derived using a variant of Dice's coefficient (van Rijsbergen, 1979) and is based upon the frequency of ⟨source, target⟩ word co-occurrence within the same example pair. *Gaijin* also incorporates a mean sentence-length bias, where ⟨source, target⟩ sentence pairs which are smaller than the overall mean length of ⟨source, target⟩ sentence pairs in the given corpus are rewarded.

The Marker Hypothesis is used to segment ⟨source, target⟩ examples. For instance, given the sentence in (51) from (Veale and Way, 1997), the Marker Hypothesis is applied. By segmenting the sentence where a categorised 'marker word' is met, the strings in (52) are produced and the associated 'marker word' is retained.

(51) In the maximum box specify the maximum amount of trap you want to add

(52) in the maximum box specify (in = Preposition)
the maximum amount (the = Determiner)
of trap (of = Preposition)
you want (you = Pronoun)
to add (to = Preposition)

The associated marker categories are then concatenated as in (53):

(53) *Preposition-Determiner-Preposition-Pronoun-Preposition*

Each ⟨source, target⟩ pair in the corpus is segmented in this way and smaller aligned fragments are derived from the chunks produced. The alignment of the sub-sentential fragments is based on word correspondence weights derived from the lexicon and similarity in ⟨source, target⟩ segment length. Furthermore, those chunks in source and target which are tagged with a similar marker category are considered more likely alignments.

In addition to each ⟨source, target⟩ marker chunk, a generalised template is produced for each ⟨source, target⟩ alignment, where well-formed constituents are replaced by a variable. For example, given the ⟨English, German⟩ sentence pair in (54), the smaller aligned fragments in (55) are produced.

(54) Displays controls for colouring the extruded surfaces ⇔
Durch Klicken auf dieses Symbol lassen sich Optionen zum Kolorieren der extrudierten Flächen anzeigen

(55) (a14) displays controls ⇔ dieses symbol lassen sich optionen
(b14) for colouring ⇔ zum kolorieren
(c14) the extruded surfaces ⇔ der extrudierten flächen

The templatised representation of these fragments is in (56).

(56) template(example-14,english,german, s(A, _, a14), s(B, prep, b14), s(C, det, c14)], durch, klicken, auf, t(A, prep, a14), t(B, prep, b14),
t(C, det, c14), anzeigen).

Where segments are replaced by a variable, a reference to the marker type of that segment is retained, e.g. (prep, det..). The templates also contain a reference to the actual chunk which has been replaced by the variable, e.g. (a14, b14, c14). The marker tags for each template are indexed in memory. For the example in (56), this would be (_,det,prep).

A similar segmentation technique is applied to an input sentence and if it has the same marked segmentation as in (53), then the associated template is retrieved and the input sentence can be matched to the source side of the template based on this structural indexing. The target segments used to produce the translation can be taken from any number of templates. We apply a similar segmentation method to (Veale and Way, 1997). However, the method in which we derive the generalised templates and word-level alignments in our system differs from that in *Gaijin*. Furthermore, we do not segment our input using the Marker Hypothesis but instead search for n-gram sequences.

2.5.1 The Marker Hypothesis as a Non-Declarative Grammar

Our system incorporates very few declarative monolingual or multilingual specifications or rules. Instead, the Marker Hypothesis is applied as a non-declarative grammar to perform a shallow parse of our ⟨source, target⟩ sentence pairs and to subsequently deduce a set of aligned chunks, words and generalised templates. Critics of our approach might suggest that the success of our methodology is dependent on the similarity of English and French. If applied to a less structurally similar language pair the system may need to be augmented with additional procedures. However, only further research would confirm if the grammar can remain essentially non-declarative.

Some procedures are currently evident within our system. For example, when aligning chunks from a given ⟨source, target⟩ example, word correspondences are firstly derived via MI (cf. section 5.3.4). For each English word, the French word with which it co-occurs most frequently is selected. However, if this word is not in the current target string then it is deemed irrelevant in that context. The next most highly co-occurring word is then selected until a corresponding target word alignment can be identified in the current target string or no alignment can be determined according to the conditions of the procedure.

Using MI allows us to determine the translations of content words in context. For example, the English word *hide* is ambiguous as it can be interpreted as a noun (*peau*) or as a verb (*masquer* or *cacher*). In (57), it is translated as the verb *masquer.* Even if the word *hide* co-occurs more frequently with the noun *peau*, the latter will not be considered as a translation in this context as it is not present in the current target string. Therefore, MI is likely to assign the word alignment in (58) to the word-level lexicon.

(57) you can also hide cell contents and formulas ⇔ vous pouvez aussi masquer le contenu et les formules des cellules

(58) hide ⇔ masquer

MI is an efficient means of deriving word translations in context. In a fully declarative rule-based system, this task would be more complex and would require the integration of large-scale rules and semantic information.

In addition, when segmenting the ⟨source, target⟩ chunks, a condition is imposed which ensures that each chunk contains at least one content word (cf. section 3.2.2). This condition, also imposed in the *Gaijin* system (Veale and Way, 1997), is useful as it can prevent the mistranslation of function and content words. For example, the English phrase *up on the roof* is currently segmented as in (59). The corresponding French phrase *sur le toit* is segmented as in (60):

(59) <PREP> up on the roof

(60) <PREP> sur le toit

They are subsequently aligned as in (61):

(61) <PREP> up on the roof ⇔ sur le toit

However, if we remove the constraint that each chunk must contain at least one function word, the same English phrase would be segmented as in (62) and the French phrase would be segmented as in (63):

(62) <PREP> up <PREP> on <DET> the roof

(63) <PREP> sur <DET> le toit

The correct alignments in (64) could potentially be deduced from the chunks in (62) and (63):

(64) <DET> the roof ⇔ the roof
<PREP> sur ⇔ on

In a case where the phrase *up on the roof* is submitted for translation, the chunks *sur* and *le toit* may be stitched together. However, the translation for the word *up* would also be retrieved from the word-level lexicon and inserted in the final translation. In all cases, this would be incorrect as a direct translation of the word *up* is omitted in this context. Ensuring that each chunk must contain a content word means that this particular error would not arise, as the chunk *up on the roof* would be retrieved from the marker-lexicon.

Although we generally view the Marker Hypothesis as non-declarative, it is possible that the aligned chunks it produces are declarative in that they could potentially be applied as 'rules' in a rule-based system. For example, the generalised template in (65) could be interpreted as a 'rule' which states that the English noun *man* can be preceded by a determiner and that its translation in French can also be preceded by a function word from the same category.

(65) <DET> man ⇔ <DET> homme

Finally, (Juola, 1994, 1998) has shown that a grammar in marker-normal form can be produced via the Marker Hypothesis (cf. p.48). One avenue for further research could be to convert our sentences into a marker-normal form grammar using the Marker Hypothesis and provide a comparison with the results derived from the methodology applied in this thesis.

2.6 Summary

In this chapter, we have provided a background to the field of MT. We have described both corpus-based and rule-based approaches to MT and we have explained our motivation for undertaking research in the EBMT paradigm. Most recent research in CBMT seems to fall under the heading of SMT. Given the benefits of EBMT, we find this surprising. We have analysed different approaches to EBMT and described how different systems have implemented the stages of matching, alignment/adaption and recombination. Many systems classified as example-based now integrate various linguistic and statistical techniques and there is little focus on EBMT for its own sake. While the notion of a hybrid system which maximises the benefits of different approaches to MT may ultimately provide the optimal solution, research within the individual MT disciplines could provide the key to this breakthrough.

We propose the development of an EBMT system which applies the Marker Hypothesis in a *linguistics-lite* approach. We have outlined the role of the Marker Hypothesis in previous NLP applications and given details of its application in MT research. Despite Juola's work, the prior research undertaken for ⟨English, French⟩ using the Marker Hypothesis is limited. In general, the application of the Marker Hypothesis to large-scale corpora has not been achieved. The METLA system only contained 30 sententially-aligned ⟨English, French⟩ pairs, while the *Gaijin* system was trained on 1,836 ⟨English, German⟩ examples. In addition, potentially useful resources such as the phrasal-lexicon (Schäler et al., 2003) and controlled language data (Carl, 2003a; Schäler et al., 2003) have not been extensively integrated in an EBMT environment, or more specifically, in an EBMT system which is based on the Marker Hypothesis.

The following chapters describe the development of our *linguistics-lite* EBMT system. We conduct experiments using phrase-based, controlled and scalable EBMT. The methodology which is applied throughout is largely based on the Marker Hypothesis and uses similar techniques to that of (Veale and Way, 1997) described in this section. Similarities can also be drawn between our work and that of (Block, 2000) (cf. section 2.4.4) and (Grefenstette, 1999) (cf. section 3.5).

Chapter 3

Phrase-Based EBMT

3.1 Motivation for a Phrase-Based Model

In chapter 2 we described rule-based and corpus-based approaches to MT. We showed that the current trend is towards corpus-based approaches and outlined our motivation for conducting this research in the EBMT paradigm.

One potential problem when developing an EBMT system is the acquisition of suitable corpora. We did not have access to a sententially-aligned corpus and therefore, despite our argument that rule-based approaches are less than successful, we introduced a novel approach where we used on-line RBMT systems to partially derive our bitext (Gough et al., 2002; Way, 2003). In chapters 4 and 5, we will describe how we seed our example-base with aligned sentence pairs. The model discussed in this chapter however, is not trained on sententially-aligned strings. Instead, the phrase is the smallest unit stored in our system's memories. There are three reasons for this:

- The translations produced by RBMT systems are less prone to error when the input strings are phrases rather than sentences;
- In contrast to TM, EBMT systems should store units smaller than sentences to facilitate the matching process;
- The advocation for the integration of the 'phrasal-lexicon' in EBMT (Schäler, 1996; Schäler et al., 2003).

RBMT systems are less likely to produce incorrect translations when they are confronted with smaller strings as there is less room for ambiguity. Consider the English sentence *A group hire lawyers to provide information about clients.* When this sentence is submitted for translation by the on-line system *Logomedia*, the translation produced is *Un avocats de la location du groupe fournir de l'information au sujet de clients.*

When translating this sentence *Logomedia* translates the verb *provide* correctly as *fournir* but fails to identify *hire* as the main verb. By translating the smaller phrases we can produce the translations in (66):

(66) a. a group ⇔ un groupe

b. hire lawyers ⇔ embaucher des avocats

c. hire lawyers to provide information ⇔ embaucher des avocats pour fournir de l'information

d. hire lawyers to provide information about clients ⇔ embaucher des avocats pour fournir de l'information au sujet de clients

e. to provide information ⇔ fournir de l'information

f. to provide information about clients ⇔ fournir de l'information au sujet de clients

g. about clients ⇔ au sujet de clients

Without any subject NP attached, the on-line system produces the infinitive form of the verb by default. Nevertheless, we can see that these smaller phrases present less ambiguity for the on-line system.

Since the initial proposal for an EBMT system (Nagao, 1984), it has been acknowledged that storing units smaller than sentences is desirable, as it more likely that new input can be matched against smaller aligned fragments.

The concept of the phrasal-lexicon (PL) was first introduced by (Becker, 1975) and has been applied successfully in different domains:

- Learnability (Zernik and Dyer, 1997);

- Text Generation (Milosavljevic et al., 1996; Hovy, 1998);

- Speech Generation (Rayner and Carter, 1997);
- Localization (Schäler, 1996).

More recently, the PL has been linked with TM technology, as a potential means of extending TMs towards fully automatic MT systems (Simard and Langlais, 2001; Schäler et al., 2003; Planas and Furuse, 2003). Given that the chances of finding exact or fuzzy matches would be greatly increased at a sub-sentential level, integrating sub-sentential alignments into the TM would potentially make more useful fragments available to the translator. EBMT systems store sub-sentential fragments along with the sententially-aligned pairs from which they are derived. Where a TM is limited to proposing fuzzy matches to a translator for him/her to adapt, an EBMT system can automatically recombine smaller fragments from the examples stored in its memories.

While some researchers have noted the potential of such a resource, few systems have exploited this knowledge to date. (Schäler et al., 2003) point out that TMs as they stand are currently being under-exploited in their potential to be developed into sophisticated EBMT systems. They propose the integration of a PL as a means of extending TMs in this direction. Figure 3.1 shows how TMs have the potential to make the transition from CAT tools to fully automatic sophisticated EBMT systems via the PL.

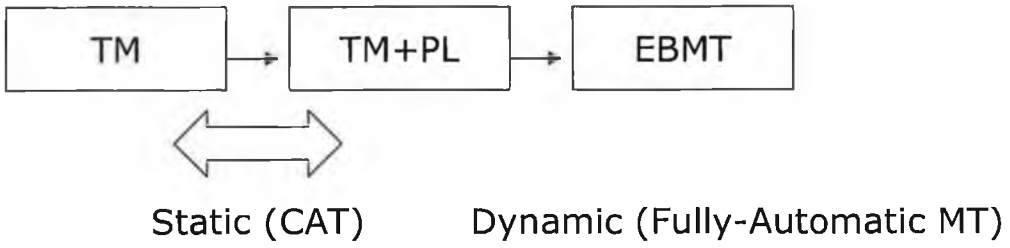


Figure 3.1: Extending TMs towards EBMT via the PL

One could foresee a situation where the current fuzzy matching techniques of TM technology are used to match phrases rather than sentences. However, recall the example in section 2.2.1, where we noted that defining a suitable threshold can often result in a trade-off between Precision and Recall. It is very likely, therefore, that any phrasal matches would be very lowly-ranked.

(Schäler et al., 2003) propose that TMs in their current sententially-aligned state possess unutilized resources. As the likelihood of finding a match with a PL is greater than at sentential level, the development of the PL based on these existing translation pairs would expand the scope of TMs and ultimately create a commercially viable EBMT system derived from an enhanced and fully exploited TM.

In this chapter, we describe how we used on-line RBMT systems to derive a set of ⟨English, French⟩ aligned phrases. Despite the fact that on-line MT is perceived to be of poor quality, we will show that we can produce reasonable results when our system's memories are seeded with the strings obtained. In section 3.2, we demonstrate how we apply the Marker Hypothesis to extract a set of additional lexical resources, i.e. chunks, words and generalised templates. We describe in section 3.3 how the translation process segments new input and retrieves chunks from our system's memories. In section 3.4 we report on a number of experiments where our system's memories were seeded with resources derived from various combinations of on-line MT systems and show how coverage and quality improve by carrying out a manual evaluation of the translations generated. As a by-product of our research, we are able to compare the performance of EBMT with the rule-based systems used to seed our example-base and provide an evaluation of these on-line systems.

A useful feature of corpus-based systems is their ability to rank translations. In section 3.4.4 we show how candidate translations are automatically weighted and ranked by the EBMT system. The problem of boundary friction is commonly associated with EBMT (cf. section 2.4.6). In section 3.5 we demonstrate how a *post hoc* validation process using the WWW can correct specific errors caused by boundary friction in the translations output by our system. Finally in section 3.6 we summarise and discuss the results obtained.

3.2 Seeding the Example-Base

3.2.1 The Phrasal-Lexicon

We developed a PL by selecting 218,697 English phrases from The Penn Treebank.[1] There are approximately 29,000 rule types in the Penn Treebank. Only 59 (0.002%) of these were used to seed our example-base. The rules selected were those instantiated more than 1000 times, while irrelevant rules such as those dealing only with numbers were eliminated. We also ensured that certain linguistic phenomena were included. For example, to ensure that the system could handle intransitive verbs, we retained rules where the Left Hand Side was a Verb Phrase (VP) and the Right Hand Side a single non-terminal.

For the 59 rule types selected, the tokens corresponding to the RHS rule were extracted. This provided an English PL. The target language translations were produced by translating the 218,697 phrases via three individual on-line MT systems. The systems chosen are listed here:

- SDL International's Enterprise Translation Server (system A)[2];
- Reverso by Softissimo (system B)[3];
- Logomedia (system C)[4].

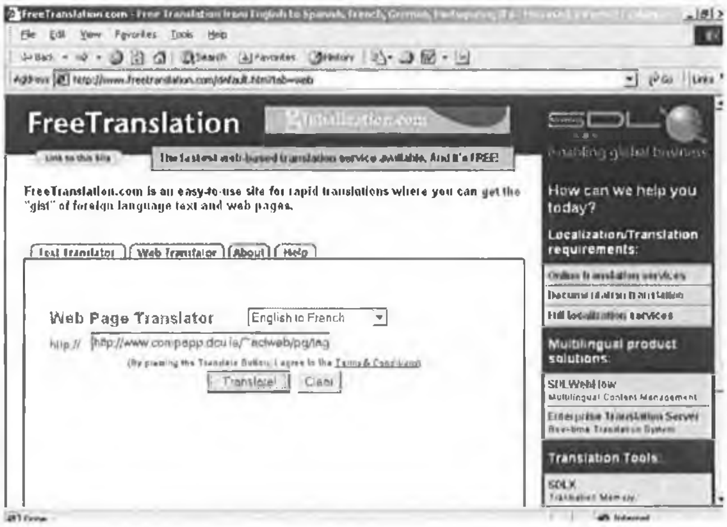
These systems were selected, not on the basis of translation quality but rather because they enabled batch translation of large quantities of text. Each English document was tagged with HTML code and sent as a web page to each system using the Unix `'wget'` function. This function takes a URL as input and writes the corresponding HTML document to a file. The translated web page retrieved is the result of a query sent in the form of a URL. Following this automatic process it is trivial to retrieve the French translations and associate them with the corresponding English source equivalents. Figure 3.2 shows how a HTML document can be submitted for translation via the on-line system *SDL*.

[1] These were extracted using a treebank tool suite (TTS) developed in Dublin City University by Aoife Cahill (http://www.computing.dcu.ie/~acahill/tts/).

[2] http://www.freetranslation.com

[3] http://www.reverso.net/text_translation.asp

[4] http://www.logomedia.net



FreeTranslation.com - Free Translation from English to Spanish, French, German, Portuguese, ...
File Edit View Favorites Tools Help
Back Search Favorites History
Address
Go Links
FreeTranslation
SDL
Enabling global business
Link to this site
Text Translator
Web Translator
About
Help
How can we help you today?
Localization/Translation requirements:
Online translation services
Documentation translation
Full localization services
Multilingual product solutions:
SDLWebFlow
Multilingual Content Management
Enterprise Translation Server
Real-time Translation System
Translation Tools:
SDLX
Translation Memory
Translate!
Clear
Done
Internet

http://www.freetranslation.com/default.htm?tab=web

The fastest web-based translation service available. And it's FREE!

FreeTranslation.com is an easy-to-use site for rapid translations where you can get the "gist" of foreign language text and web pages.

Web Page Translator English to French

http:// http://www.compapp.dcu.ie/~acweb/pg/eg

(By pressing the Translate Button, I agree to the Terms & Conditions)

Figure 3.2: Submitting a HTML-tagged file for translation via on-line system *SDL*

Assuming the file submitted contains a single tagged phrase, *<ul>the problem<\ul>*, an output file is retrieved in HTML format as *<ul>le problème<\ul>*. When the tags are stripped from the output file, we can associate the translations with their original source strings given that the line numbers in the source and target files correspond.

When translating VPs without an attached NP subject, the on-line systems produce the infinitive form of the verb by default. In order to obtain the finite form of the verb, we attached dummy subjects. Initially these were third person plural pronouns, causing translations in the same form to be produced. Later we also included third person singular pronouns to decrease the bias of the system towards sentences which contained pronouns in third person plural form. We experiment using these verb forms separately to seed the memories of our EBMT system and also by combining both singular and plural fragments in the system's memories.

3.2.2 The Marker-Lexicon

As in the approach of (Veale and Way, 1997; Juola, 1997) (cf. section 2.5), we apply the Marker Hypothesis in a pre-processing stage to segment the aligned ⟨source, target⟩ phrases. For the source and target languages (English and French respectively), we exploit a set of known marker words to indicate the beginning and end of segments. We use seven categories in total. This closed list comprises determiners (<DET>), prepositions (<PREP>), quantifiers (<QUANT>), conjunctions (<CONJ>), possessive pronouns (<POSS>), personal pronouns (<PPRON>) and wh-adverbs (<WRB>). For English we use the set of marker words in (67):

(67) <DET> the, a, an, those, these,...
<PREP> in, on, out, with, from, to, under,...
<QUANT> all, some, few, many,...
<CONJ> and, or...
<POSS> my, your, our...
<PPRON> i, you, he, she, it..
<WRB> when, what...

For French, we use the a similar set listed in (68):

(68) <DET> le, la, l', les, ce, ces, ceux, cet...
<PREP> dans, sur, avec, de, à, sous...
<QUANT> tous, tout, toutes, certain, quelques, beaucoup...
<CONJ> et, ou,...
<POSS> mon, ma, mes, ton, ta, tes, notre, nos,...
<PPRON> je, j', tu, il, elle,...
<WRB> quand, quelle, quel, quelles...

In a pre-processing stage, we traverse the ⟨source, target⟩ phrases word by word. Whenever a marker word is encountered, this signals the beginning of a new chunk which is labelled with its marker category. (69) illustrates the results of running the Marker Hypothesis over the English phrase *of his duties and prerogatives.*

(69) <PREP> of <POSS> his duties <CONJ> and prerogatives

As in the *Gaijin* system (Veale and Way, 1997), we impose a further constraint which ensures that each chunk must contain at least one non-marker word. For example, the underlined string *of his duties* will be retained as a single chunk, labelled with <PREP>. This prevents the formation of a chunk such as *of his*, which is devoid of content words.

For all of the 218,697 English phrases extracted from the Penn Treebank, we obtain three translations (one from each of the on-line systems A, B and C, p.57). For example, for the English phrase *of his duties and prerogatives*, systems A, B and C produce the translations in (70):

(70) a. de ses devoirs et de prérogatives

b. de ses impôts et prérogatives

c. de ses devoirs et prérogatives

We apply the Marker Hypothesis to each set of ⟨source, target⟩ translations, segmenting the phrases as described above. We derive our marker-lexicons on the naïve yet effective assumption that marker-headed chunks in the source S map sequentially to their target equivalents T, subject to their marker categories matching. The translation produced by on-line system C for *for his duties and prerogatives* was *de ses devoirs et prérogatives.* Applying the Marker Hypothesis to the French phrase gives us the segmented phrase in (71):

(71) <PREP> de <POSS> ses devoirs <CONJ> et prérogatives

As with the source sentence in (69), an additional caveat ensures that each chunk contains a content word. From here we can derive the marker chunks in (72). In this way, from the original aligned ⟨source, target⟩ phrase we can derive even smaller aligned fragments and consequently increase the potential that new input can be matched against the examples in our system's memories.

(72) *System C*: <PREP> of his duties ⇔ de ses devoirs
<CONJ> and prerogatives ⇔ et prérogatives

The target strings derived from on-line systems A and B are also segmented and using the same methodology the chunks are aligned with those in (69). As a result, the aligned chunks in (73) are also produced:

(73) *System A*: <PREP> of his duties ⇔ de ses devoirs
<CONJ> and prerogatives ⇔ et de prérogatives

System B: <PREP> of his duties ⇔ de ses impôts
<CONJ> and prerogatives ⇔ et prérogatives

3.2.3 The Word-level Lexicon

We can derive word alignments from marker chunks such as those in (72) and (73). We base our word alignments on the assumption that where a ⟨source, target⟩ chunk contains just one non-marker word in both source and target, these words are translations of each other. From the example in (72) therefore, we can derive the word alignments in (74):

(74) <CONJ> and ⇔ et
<LEX> prerogatives ⇔ prérogatives

In this way, we can extract smaller aligned segments from the phrasal-lexicon without recourse to complex parsing techniques. Any content words derived in this manner are assigned a <LEX> tag and added to the word-level lexicon. (Juola, 1994, 1997) tags words ending in '-ed' as verbs. We do not mark verbs in our approach, as they are not considered a closed class. However, as English strings were derived from rules in the Penn Treebank, we can assume that the phrasal chunks correspond to the rule RHS. For example, a rule in the Penn Treebank such as VP ⇔ VBG, NP, PP indicates that the first word in each string corresponding to the RHS of the VP rule is a VBG (present participle). Such words are also marked with a <LEX> tag.

3.2.4 The Generalized lexicon

In a final pre-processing stage we create a set of generalised templates when we replace marker words with their associated marker tag. This is similar to the process of (Block, 2000) (cf. section 2.4.4). Consider the example in (75), contained in our phrasal-lexicon.

(75) <CONJ> and prerogatives ⇔ et prérogatives

If our system was confronted with the input *or prerogatives*, it would not be able to translate this phrase, assuming that it is not present in the marker-lexicon. However, by generalising over the example in (75), the template in (76) is produced by replacing the lexical items *and* and *et* in the source and target phrases, with their marker tag, <CONJ>:

(76) <CONJ> prerogatives ⇔ <CONJ> prérogatives

The input phrase *or prerogatives* is then converted to its generalised form, <CONJ> *prerogatives*, and can now be successfully matched against the source side of the template in (76). From the word-level lexicon we can retrieve the entry '<CONJ> or : ou' (assuming that this exists), and the translation can be inserted into the target side of the template, producing the final string, *ou prérogatives.*

In this way, we increase the coverage of our system by clustering on marker words and creating more flexible templates to facilitate the matching process.

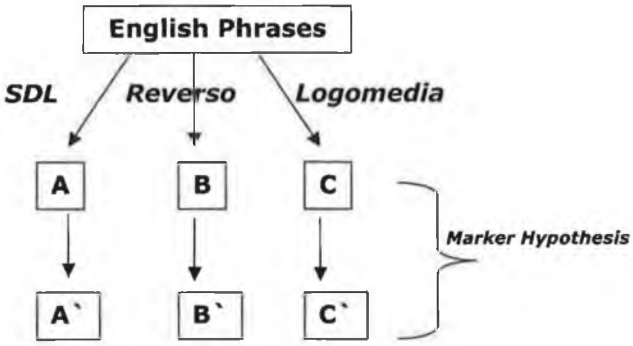


Figure 3.3: Summary of the Knowledge Sources in the phrase-based EBMT system

3.2.5 Summary of Knowledge Sources

For each English phrase extracted from the Penn Treebank (218,697 in total), a translation is produced by three different on-line MT systems: *SDL's Enterprise Translation Server* (system A), *Reverso* by *Softissimo* (system B) and *Logomedia* (system C). When translating VPs we attach dummy subjects in the form of both singular and plural pronouns to obtain finite verb forms. We then apply the Marker Hypothesis to each set of 218,697 ⟨English, French⟩ phrases in order to segment them and produce even smaller aligned fragments. This provides us with three additional knowledge sources A′, B′ and C′ and gives us a total of 656,901 ⟨source, target⟩ translation pairs. From here, we generalise our alignments by replacing certain marker words with their associated tag and as a by-product of this process, we derive a word-level lexicon. Figure 3.3 illustrates the knowledge sources which seed the memories of our EBMT system. A′, B′ and C′ represent all lexicons derived via the Marker Hypothesis i.e. the marker-lexicon, the word-level lexicon and the generalised-lexicon. In the following sections we will show how these knowledge sources can be used to derive translations for novel NPs and sentences and give detailed experimental results.

3.3 Retrieving chunks and Performing Translation

In our EBMT system, we apply an n-gram-based segmentation method. Originally, we segmented our input sentences into all possible bigrams, trigrams etc. and performed a search for these strings within our system's memories. However, the manner in which the Marker Hypothesis is applied to create our marker-lexicons means that some of these n-grams will never be found. For example, when segmenting the string *the decline at the end of the year*, a new chunk would begin each time a marker word is encountered, producing the chunks in (77):

(77) <DET> the decline <PREP> at the end <DET> of the year

Therefore, n-grams such as *the decline at*, *the end of* or *the end of the* will never be located within our lexicons. Taking this into consideration and excluding those n-grams which end in a marker word, we search our lexicons for all remaining bigrams, trigrams etc. The lexicons which seed the memories of our EBMT system are created by deriving smaller aligned fragments and applying a generalisation process (cf. section 3.2.4). For each n-gram, the system's memories are searched from maximal context (specific ⟨source, target⟩ sentence/phrasal pairs) to minimal context (word-level correspondences). The order in which the resources are searched is:

- original bi-text (⟨source, target⟩ phrasal alignments) (cf. section 3.2.1);
- marker-aligned chunks (cf. section 3.2.2);
- generalised marker chunks(cf. section 3.2.4);
- word-level lexicon (cf. section 3.2.3).

Each chunk in the system's lexicons is stored along with all its possible translations.[5] When an *n*-gram sequence is located, its candidate translations are retrieved. For example, if the system is confronted with the English sentence in (78), it is segmented into the *n*-grams in (79). Given that our lexicon is phrase-based, some of these strings will not be found e.g. *sales are booming.* However, in order to make our segmentation method more portable to other marker-based models (cf. chapters 4 and 5), we only eliminate a search for those *n*-grams which end in a marker word as described above.

(78) The monthly sales are booming

(79) The monthly sales are booming (5-words/exact match)
the monthly sales are, monthly sales are booming (4-words)
the monthly sales, monthly sales are, sales are booming (3-words)
the monthly, monthly sales, sales are, are booming (2-words)
the, monthly, sales, are, booming (1-word)

Initially, the phrasal-lexicon is searched. Assuming that an exact match is not found (in this case we know that this string will not be found as we do not store any sententially-aligned strings in our example-base), the marker-level lexicon is searched in an effort to locate any *n*-grams from this knowledge source.

When the system's memories are seeded with strings derived from on-line system C, the *n*-gram in (80) is retrieved from our marker-lexicon.

[5] The sub-sentential alignment algorithm can produce multiple translations for a single chunk, depending on the number of times the chunk occurs within the corpus. Some of these alignments may be better than others.

(80) are booming ⇔ prospère
are booming ⇔ prospèrent

Any *n*-grams which are not located are generalised and the generalised-lexicon is searched. In this case the *n*-gram in (81) is located in our generalised-lexicon.

(81) <DET> monthly sales ⇔ <DET> ventes mensuelles

Finally, if any words have not been located, the word-level lexicon is searched in order to locate these. Our word-level lexicon returns the translations of *the* in (82):

(82) the ⇔ le
the ⇔ l'
the ⇔ les
the ⇔ la

The French translations of the word *the* in (82) can be inserted into the target side of the generalised template in (81) to generate the strings in (83).[6]

(83) les ventes mensuelles
l' ventes mensuelles
la ventes mensuelles
le ventes mensuelles

A set of candidate translations for the sentence in (78) can be produced by combining the strings in (83) with those in (80). This would result in the generation of the translations in (84).

[6] An integrated weighting process means that in reality only one of these strings will be produced in this instance (cf. 3.4.4).

(84) a. les ventes mensuelles prospère

b. les ventes mensuelles prospèrent

c. l' ventes mensuelles prospère

d. l' ventes mensuelles prospèrent

e. le ventes mensuelles prospère

f. le ventes mensuelles prospèrent

g. la ventes mensuelles prospère

h. la ventes mensuelles prospèrent

The translations in (84) are produced by recombining the translations retrieved for each chunk. Rather than generate all possible ordering of chunks and words, as English and French are both Subject-Verb-Object (SVO) languages, we base the ordering of the target language chunks on the source language sentence. Of course, this assumption is not extensible to all other language pairs and in chapter 6 we provide some alternative solutions. Nevertheless, for the purpose of ⟨English, French⟩ translation it is generally sufficient. We envisage that the sub-sentential chunks retrieved from our system's memories will contain enough context to prevent many instances of boundary friction or incorrect word order.

Of course, as is evident from the translations in (84), our system, as is the case with the majority of EBMT systems, still suffers from boundary friction and only the translation in (84b) is syntactically correct. It displays agreement between the plural determiner *les* and the plural noun *ventes* and noun-verb agreement between *ventes* and the third person plural verb *prospèrent.* In section 3.4.4, we show how these translations can be ranked and output with associated weights to facilitate the pruning of a list of candidate translations. In section 3.5 we illustrate how the identification of the 'best' translation can be further assisted by a *post hoc* validation process and show how certain cases of boundary friction can be significantly reduced. In the following section we outline some experiments designed to test our system and comment on the results obtained.

3.4 Experiments and Results

Our experiments involved translating two test sets of data. Initially, we translated a set of 500 Noun Phrases (NPs). Our reasons for doing so were twofold. Firstly, we wanted to ensure that nominal phrases were being translated correctly by the system and secondly, we wanted to identify cases where our system could improve upon the translations derived from the on-line systems. We also translated 200 sentences. Examples of both test sets are in (85):

(85) *Noun Phrases*

- the flexibility of private voluntary organisations
- this Japanese investment in the mechanical engineering industry
- an increase through issues of new shares and convertible bonds

Sentences

- A major concern for the parent company is what advertisers are paying per page
- his empire owed the steel company a 14% increase through issues of new shares and convertible bonds
- the global trade issues complicate a valuation of the new company

We created the test sets automatically by combining words and phrases from the Penn-II Treebank. We ensured that the strings in the test set reflected the frequency of their corresponding rule type, and also that each of the 59 rule types were included in the test set.

We performed a number of experiments to test the coverage and quality of translations produced by our EBMT system. We seeded the example-base with source strings and translations derived from each of the individual on-line MT systems. We then used various combinations of translations derived from the different systems:

- Translations derived from each of the individual on-line MT systems, A, B and C;

- Translations derived from each pair of different on-line MT systems, AB, AC and BC;

- Translations derived from all three on-line MT systems, ABC.

In order to coax our system into producing verb forms other than the default infinitive, we attached dummy subjects before submitting the phrases for translation by the on-line systems. We compared and contrasted the results obtained when the memories of our system were seeded with the source strings and the translations derived using both third person singular and third person plural dummy subjects.

We perform a manual evaluation on the translations produced by our system, measuring coverage and quality. Coverage refers to the number of translations which the system can produce. Quality is measured using a scale of 1-3 as listed below[7]:

- *Score 1*: Contains major syntactic errors and is unintelligible;

- *Score 2*: Contains minor syntactic errors and is intelligible;

- *Score 3*: Contains no syntactic errors and is intelligible.

As a result of the manual evaluation using these figures, we can assess the effects of using single and multiple on-line systems on translation quality. In an effort to identify where our EBMT system can improve on the translations output by the rule-based systems, we translated the sentences from our test set directly using the on-line MT systems. As a result, we can assess the overall gain of using EBMT over rule-based MT and can provide some insights into the quality of translations produced by the on-line systems.

3.4.1 Experiments using Single Knowledge Sources: Sentences

When translating the initial source language phrases via the on-line systems, we used third person singular and third person plural dummy subjects, thus obtaining both verb forms in the target language. When translating our 200 sentence test set we seeded the system's memories with translations derived using individual on-line systems (A, B and C). At different stages, we used translations derived from the application of third person

[7] This scale is intuitive.

plural, third person singular and both third person plural and third person singular dummy subjects. The average length of a sentence in our test set is 8.5 words. The minimum and maximum lengths of the sentences were 3 words and 18 words respectively.

Experiments using Third Person Plural Subjects

Our EBMT system obtained 92% coverage when its memories were seeded with chunks derived from system A *Enterprise Translation Server* and system C *Logomedia* (184/200 sentences were translated in each case). When the chunks used to seed our EBMT system were derived from System B *Reverso*, we obtained a slightly lower coverage of 90% (180/200 sentences were translated). The reason that some sentences are not translated is due to the absence of a word in the word-level lexicon. For example, given the source sentence in (86), the chunks in (87) are located in the system's marker-lexicon.

(86) Those investments raised the initial transaction

(87) those investments ⇔ ces investissements
the initial transaction ⇔ la transaction initiale

The word *raised* is located in the system's memories but only as a past participle (88), and therefore cannot be matched against the verb in the input sentence which is in simple past form.

(88) were raised ⇔ ont été élevés

For cases such as this, we insert the untranslated source word, producing the string in (89). However, we do not consider this to be a complete translation.

(89) those investments raised the initial transaction ⇔ ces investissements *raised* la transaction initiale

Given the manner in which the initial source phrases were translated (attaching third person plural dummy subjects), lower quality translations are produced for sentences which contain verbs in the third person singular. For example, when translating the sentence in

(90), we obtain the chunks in (91) when the system's memories are seeded with translations derived from on-line system A.

(90) The girl reports an increase in the salary at the end of the trading session

(91) the girl ⇔ la fille
reports ⇔ rapportent
an increase ⇔ une augmentation
in the salary ⇔ dans la salaire
at the end of the trading session ⇔ à la fin de la session du commerce

Of course, when these fragments are recombined, conflict arises between the singular third person NP *La fille* and the third person plural verb *rapportent*. In section 3.5 we demonstrate how such problems of NP-VP boundary friction may be overcome *post hoc*. However, in the next section, we will also demonstrate that seeding the example-base with additional strings using a combination of verb forms can also serve to alleviate this problem. Table 3.1 shows the translation quality for 200 sentences, when chunks are derived from individual on-line MT systems with third person plural (3PL) dummy subjects attached in the derivation of the initial fragments used to seed the example-base.

System	*Score 1*	*Score 2*	*Score 3*
A - *SDL*	14.2%	51.2%	34.6%
B - *Reverso*	8.9%	54.7%	36.4%
C - *Logomedia*	4.4%	59.1%	36.5%

Table 3.1: Translation Quality for Sentences: chunks derived from individual on-line MT systems, 3PL dummy subjects

We observe that seeding the system's memories with strings derived from on-line system *Logomedia* (system C) yields marginally better translation results. With regard to score 3, System C outperforms system A by 1.9% and system B by 1.1%. However, when all intelligible translations are taken into account (Score 2 and 3), System C outperforms system B by 5.5% and system A by 9.8% These initial results suggest that *Logomedia* might be the better on-line system. However, as yet the difference between the systems is too small to substantiate such an hypothesis. In section 3.4.6, we will provide further evidence in favour of *Logomedia* being the better system.

Seeding the Database with more Examples

In the previous section, we showed how seeding the memories of the system with translations derived using third person plural dummy subjects biased the system in favour of third person plural NPs and reduced the translation quality for sentences containing third person singular NPs. When the target strings are produced using third person singular dummy subjects the bias is reversed. Figure 3.4 shows the performance of our EBMT system when its memories are seeded using the individual on-line MT systems and attaching dummy subjects in both singular and plural form.

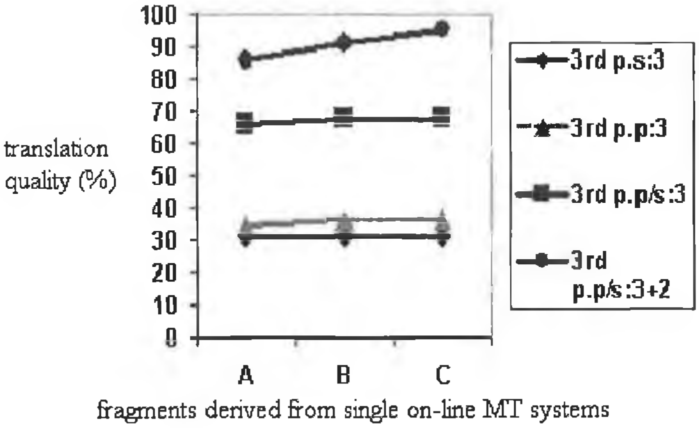


Figure 3.4: Translation quality when our system's memories are seeded with strings derived from individual on-line MT systems and using both 3PL and 3PS dummy subjects

Given that there is a larger number of NPs in third person plural form in our test set, the system's performance is slightly lower when just third person singular (3PS) dummy subjects are used. The number of translations which receive a high score of 3 for quality deteriorates for all systems (approximately 5% for system A and 3% for systems B and C).

However, by using the fragments derived from both singular and plural third person dummy subjects, we noted a considerable improvement in translation performance. 66.1% of translations produced when the system's memories are seeded with chunks derived from system A are rated 3. When chunks from system B are used, this figure is 69.7% and for system C, 68%. When we consider intelligible translations, i.e. those rated 2 or 3, the scores are as follows: System A, 85.8%, system B, 91.1% and system C, 95.6%. This is further evidence that *Logomedia* is probably the best system.

3.4.2 Experiments using Single Knowledge Sources: Noun Phrases

We translated 500 NPs. The average length of these NPs was 6.14 words, the maximum length was 12 words and the minimum length 3 words. Similar to the sentence test set, when using the fragments derived from systems A and C, a higher coverage of 94.8% of NPs (474/500) is obtained than when system B is used. When the latter is used to derive the fragments in our example-base, 92.6% of NPs (463/500) are translated.

The coverage and quality of the translations produced for NPs when the individual on-line systems are used to seed the memories of our system is shown in Table 3.2.

System	*Coverage*			*Quality*	
			Score 1	**Score 2**	**Score 3**
A - *SDL*	94.8%		13.7%	52.5%	33.8%
B - *Reverso*	92.6%		10.6%	52.3%	37.1%
C - *Logomedia*	94.8%		4%	48.7%	47.3%

Table 3.2: Translation Coverage and Quality for NPs: chunks derived from individual on-line MT systems

In order to produce a translation, we break each NP into *n*-grams as described in section 3.4.4 and search the phrasal and marker-lexicons for matching sequences. If a translation cannot be produced from the *n*-grams located, we search the generalised-lexicon and word-level lexicons for relevant template and word matches.

In the cases where a translation cannot be produced for an NP, this is mainly due to the absence of a relevant generalised template. For example, when translating the input NP *my high test scores*, we fail to find a match in the phrasal or marker-lexicons. The next step is to create the generalised form of the NP as in (92):

(92) <POSS> high test scores

Given that the closest matching template is <DET> *high test scores*, we cannot retrieve the associated translation as the marker tags <DET> and <POSS> are not the same. However, this constraint is not binding and it may be possible to revert to an even more general category where <DET> and <POSS> tags are considered similar.

3.4.3 Experiments using Multiple Knowledge Sources

In another experiment, we used various combinations of the fragments derived from the individual on-line MT systems to seed our example-base. We use translations derived using pairs of on-line systems (AB, AC and BC) and a combination of all three on-line systems (ABC). As we increase the number of fragments in our system's memories, we increase the number of chunks retrieved and ultimately the number of translations produced. For example, Table 3.3 shows the number of translations produced for the NP *a plan for reducing debt over 20 years*, when our system's memories are seeded with translations derived using the individual on-line systems.

System	*No. Translations*
A	14
B	10
C	5

Table 3.3: No. of translations produced for the NP *a plan for reducing debt over 20 years* when our system's memories are seeded with chunks derived from individual on-line systems

When chunks derived from multiple on-line systems are used to seed our example-base, the number of translations produced for the same NP increases as illustrated in Table 3.4.

System	*No. Translations*
AB	108
AC	72
BC	42
ABC	224

Table 3.4: No. of translations produced for the NP *a plan for reducing debt over 20 years* when our system's memories are seeded with chunks derived from multiple on-line systems

As the on-line systems can produce different translations, there are more potential chunk translations available in the system's memories and therefore more translations can be produced. In the following section, we will show how using a combination of on-line systems to seed our example-base not only increases the number of translations produced by our EBMT system for each input sentence but also improves the quality of the translations output by the system.

Combining Fragments from Different On-Line Systems : Sentences

The 16 sentences that were untranslated when individual systems A and C were used to seed the example-base remain untranslated when a combination of the translations produced by all three systems (ABC) is used. However, there is a significant improvement in the quality of the translations generated when the chunks from multiple on-line systems are combined. This is illustrated in Figure 3.5:

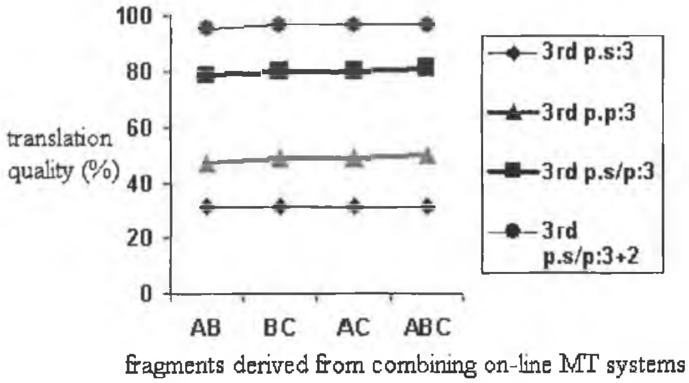


Figure 3.5: Translation quality when the system's memories are seeded with chunks derived from multiple on-line MT systems and 3PS and 3PL dummy subjects are used

Previously, when single on-line systems were used to seed the example-base (cf. section 3.4.1), the best performance was 36.5% of translations obtaining score 3, i.e. translations were produced which contained no syntactic errors and were intelligible. When a combination of the two systems is used to seed the memories of our example-base, this figure rises to 48.9%. When resources from all three on-line systems are combined, 50% of translations output by our EBMT system obtain a high score of 3.

Combining Fragments from Different On-Line Systems and Seeding the Example-Base with Additional Examples: Sentences

Further improvements in translation quality can be seen when the system is seeded with additional examples, i.e. translations were derived using both third person singular and third person plural dummy subjects. The best pairwise performance now rises to 80.4%.

When all three on-line systems are used (ABC), 81.5% of translations now obtain a score of 3. When we consider intelligibility, i.e. those translations with a score of either 2 or 3, all combinations apart from AB produce 96.7% intelligible translations. 96.5% of translations produced by AB are considered intelligible.

Combining Fragments from Different On-Line Systems: NPs

Table 3.5 shows the effects of combining resources on the quality and coverage of translations obtained for our NP test set.

Combinations	*Coverage*	*Quality (score 3)*
A	94.8%	33.8%
B	92.6%	37.1%
C	94.8%	47.3%
AB	95.4%	54.1%
BC	95.6%	64.0%
AC	94.8%	72.0%
ABC	96.0%	77.8%

Table 3.5: Translation Coverage and Quality for NP's: chunks derived from different combinations of on-line MT systems

When individual on-line systems are used to seed the example-base, the best performance for NPs is a score 3 for 47.3% of translations (system C). The worst performance (system A) is a score 3 for 33.8% of translations. When resources from different on-line systems are combined and the NPs are translated using the fragments from different systems, translation quality rises considerably. When all three on-line systems are used to seed the example-base, 77.8% of NP translations obtain a score of 3. To give an example of how combining fragments from different systems leads to improved translation performance, consider the input phrase in (93):

(93) an even bigger bundle in common

When the system's memories are seeded with chunks derived from on-line system B (*Reverso*) the translation in (94) is produced:

(94) un paquet même plus grand dans commun

We can see from this example that *Reverso* incorrectly translates the prepositional phrase *in common* as *dans commun.* When the phrase is translated using chunks derived from on-line system C (*Logomedia*), the translation in (95) is produced:

(95) un même plus grand paquet en commun

The prepositional phrase *in common* is correctly translated as *en commun.* However, the translation of the noun phrase *an even bigger packet* is erroneous as the word order is incorrect. Nevertheless, using a combination of resources derived from systems C (*Logomedia*) and B to seed our system, the improved translation in (96) is derived among others:

(96) un paquet même plus grand en commun

3.4.4 Producing Weights and Ranking Translations

One aspect of corpus-based systems is their ability to output many different translations. We output each candidate translation produced by our system with an associated weight and rank the translations according to these weights. In the following section, we describe how these weights are calculated.

Calculation of Weights

When the system is confronted with the input sentence *the boy went to the river*, it is segmented into n-grams as described in section 3.3. The associated translations in (97) were located in the system's memories.[8] These were retrieved along with the number of times each translation exists within the system's memories.

(97) the boy ⇔ le garçon (5)
went to the river ⇔ allait au fleuve (3)
went to the river ⇔ allait à la rivière (2)
went to the river ⇔ allaient au fleuve (3)
went to the river ⇔ allaient à la rivière (2)

The translations of the relevant source language chunks are recombined to produce a set of candidate translations. Each translation produced has an associated weight and

[8] The list of retrieved chunks has been shortened for the purpose of this example.

these weights are used to rank the translations. This figure is calculated using the formula in (98):

(98) $$weight = \frac{no.\ occurrences\ of\ the\ proposed\ translation}{total\ no.\ of\ translations\ produced\ for\ source\ language\ chunk}$$

Producing a set of weighted translations is a common and useful feature of probabilistic systems. While some incorrect translations may be produced, the correct translation may also be generated by the system. It is hoped that by ranking the translations using this technique, the higher weighted translations will also be 'better' translations than those ranked lower down. Any user of the system, therefore, will not have to sift through hundreds or even thousands of translations to identify the correct one.

For the source language chunk *went to the river*, four translations are retrieved.[9] According to the formula in (98), the weight for these chunks can be calculated as:

- P(went to the river | allait au fleuve) (3/10)
- P(went to the river | allait à la rivière) (2/10)
- P(went to the river | allaient au fleuve) (3/10)
- P(went to the river | allaient à la rivière) (2/10)

The source language chunk *the boy* has a unique translation *le garçon*. This chunk pair occurs five times in our example-base and therefore its weight is calculated as:

- P(the boy | le garçon) (5/5)

The weight for the final translation is calculated by multiplying the weights for each individual chunk comprising the final translation. The candidate translations produced from the above example are in (99):

[9] These translations are retrieved from the system's memories when they are seeded with resources from all three on-line systems and both singular and plural dummy subjects are used.

(99) a. Le garçon allait au fleuve $(5/5)*(3/10)=0.3$

b. Le garçon allaient au fleuve $(5/5)*(3/10)=0.3$

c. Le garçon allait à la rivière $(5/5)*(2/10)=0.2$

d. Le garçon allaient à la rivière $(5/5)*(2/10)=0.2$

The number of on-line systems used to seed our system's memories is also factored into the weighting process. In an initial experiment (cf. section 3.4.1), translations derived from individual on-line systems (A, B and C) were used to seed the system's memories. The Marker Hypothesis was then applied to these aligned pairs to deduce smaller aligned fragments. From each on-line system, therefore, we can derive two knowledge sources or six in total: (A and A′, B and B′ and C and C′). When we seed our example-base with translations derived from a single on-line system we add the weights of the translations produced and divide by 2. If pairs of on-line systems are used to seed the example-base, the number of knowledge sources rises to four so we divide the weights produced by 4. Similarly, using all three on-line systems to seed our example-base means that our system has recourse to six lexical resources and therefore we divide the weights by 6.

The above translations were produced when the system's memories were seeded with chunks derived from all three on-line systems (ABC) and therefore the final weight for each translation is arrived at by dividing the figures in (99) by 6.[10] The final weights are calculated as:

(100) a. Le garçon allait au fleuve $(((5/5)*(3/10))/6)=0.05$

b. Le garçon allaient au fleuve $(((5/5)*(3/10))/6)=0.05$

c. Le garçon allait à la rivière $(((5/5)*(2/10))/6)=0.0333333$

d. Le garçon allaient à la rivière $(((5/5)*(2/10))/6)=0.0333333$

We can see that the translations in (100a) and (100b) are ranked higher than those in (100c) and (100d). This is because the translations of the chunk *went to the river* occurs more frequently as *allait au fleuve* and *allaient au fleuve* than *allait à la rivière*

[10] Note that the weights in (99) reflect the likelihood of a translation t given a source string s and that the figures sum to 1. The weights in (100) reflect the weights over the entire corpus. However, the relative probabilities and ranking remain unaltered.

and *allaient à la rivère*.[11] Of course, the translations in (100a) and (100c) are well-formed as the singular subject NP *le garçon* agrees with the singular past participle in the VP *allait*. The translations in (100b) and (100d) on the other hand are ungrammatical as they contain a third person plural past participle *allaient*. Ideally we would like our system to rank the translations in (100a) and (100c) higher than those in (100b) and (100d). In this case, the correct translation (100a) and the ungrammatical translation (100b) are jointly ranked first. In section 3.5, we will show how a *post hoc* validation process can identify the better translation and raise the status of translation (100a) in terms of where it is ranked by the system. In the following section we will show how this ranking process can successfully place the 'best' translation within the top 1% of translation results produced by our system.

Ranking: Sentences

A human expert was assigned the task of locating the 'best' translation output by our system for each sentence. We then identified where our ranking mechanism correlated with human judgement. Table 3.6 shows where the 'best' translation as identified by a human was ranked by our system when individual on-line systems were used to seed the system's memories. In over 65% of cases, the 'best' translation was also ranked first by the system. In each case, the 'best' translation was not located outside the top five automatically-ranked translations.

System	*Ranked1*	*Ranked 2-5*
A	71.6%	28.4%
B	65.3%	34.7%
C	70.3%	29.7%

Table 3.6: Ranking of 'Best' Translations for Sentences: chunks derived from individual on-line MT systems using 3PL dummy subjects

When translations derived from both third person singular and plural dummy subjects are used to seed the example-base, there is a slight deterioration in the number of 'best' sentences which are ranked first by the system. This is illustrated in Table 3.7.

[11] Only third person singular and third person plural forms of the verb exist within the example-base. Attaching alternative dummy subjects would provide us with different verb forms.

System	*Ranked 1*	*Ranked 2-5*	*Ranked 6-10*	*Ranked 10-20*
A	65.2%	30.5%	0%	4.3%
B	60.8%	34.9%	0%	4.3%
C	64.1%	31.6%	0%	4.3%

Table 3.7: Ranking of 'Best' Translations for Sentences: chunks derived from individual on-line MT systems using 3PL + 3PS dummy subjects

We see that for systems A and C, the number of translations considered 'best' and also ranked first decreases by approximately 6%, and for system B by 4.5%. In addition, the 'best' translation now occurs within the top 20 automatically-ranked translations, as opposed to within the top five translations when only strings derived from third person plural dummy subjects are used to seed the example-base. Increasing the example-base, therefore, directly affects the ranking process.

Table 3.8 shows where the 'best' translation was located in the set of ranked translations output by the EBMT system when multiple knowledge sources are used to seed the example-base.

System	*Ranked 1*	*Ranked 2-5*	*Ranked 6-10*
AB	67.6%	31.1%	1.3%
AC	54%	46%	0%
BC	63.6%	35.1%	1.3%
ABC	62.2%	35.1%	2.7%

Table 3.8: Ranking of 'Best' Translations for Sentences: chunks derived from multiple on-line MT systems using 3PL dummy subjects

We note that although the number of translations output by the system increases considerably, the 'best' candidate can be located within the top five translations output by the system in over 97% of cases and invariably within the top ten translations. In at least 54% of cases, the 'best' translation is also ranked first by the system.

We predicted that using multiple on-line systems to seed the example-base and including strings derived from both singular and plural dummy subjects would also affect the ranking process. Table 3.9 shows the ranking of the 'best' translation in this case.

We can see that the number of cases where the 'best' translation is ranked first by the system decreases for each combination. For AB we note a 24% decrease, for AC 15% and for BC 20%. When all systems are used to populate our example-base and both third

System	*Ranked 1*	*Ranked 2-5*	*Ranked 6-10*	*Ranked 10-20*	*Ranked 20-40*
AB	43.4%	32.6%	2.3%	13%	8.7%
AC	39.1%	34.8%	5.4%	12%	8.7%
BC	43.4%	31.7%	2.7%	13.5%	8.7%
ABC	35.2%	28%	9.4%	18.3%	9.1%

Table 3.9: Ranking of 'Best' Translations for Sentences: chunks derived from multiple on-line MT systems using 3PL + 3PS dummy subjects

person singular and plural dummy subjects are used (ABC), we note a decrease of 27% in the number of 'best' translations ranked first. The 'best' translation is sometimes found as low as 36th, whereas previously it could be located in the top ten. However, we can still find the 'best' translation within the top five in over 63% of cases and in the top ten 72.6% of the time. Moreover, the 'best' translation can still be located by examining the top 1% of translations output by the system despite the fact that for some source language sentences more than 2000 candidate translations are produced.

Ranking: NPs

When individual on-line systems are used to seed our example-base, the 'best' translation is ranked within the top ten translations output by our system for our NP test set. Table 3.10 shows where our 'best' translation was ranked by the system:

System	*Ranked 1*	*Ranked 2*	*Ranked 3-5*	*Ranked 6-10*
A	64.6%	9.1%	23.6%	2.7%
B	57.7%	15.6%	24.8%	1.9%
C	60%	7.6%	29.3%	3.1%

Table 3.10: Ranking of 'Best' Translations for NPs: chunks derived from individual on-line MT systems

While all 'best' translations are ranked within the top ten, the vast majority of 'best' translations output by the system (96%) are ranked in the top five. The best translation is ranked first over 57% of the time.

When fragments from multiple on-line systems are used to seed the example-base, the best translation output by the system remains in the top ten. This means that any potential user of the system would only need to search through the top 1% of translation candidates to identify the 'best' translation, thus facilitating the pruning of translations output by the system.

The ranking of NP translations when multiple on-line systems are used to seed the example-base is summarised in Table 3.11.

System	*Ranked 1*	*Ranked 2*	*Ranked 3-5*	*Ranked 6-10*
AB	42.2%	13.8%	41.3%	2.7%
AC	62.1%	14.1%	21.3%	2.5%
BC	66.4%	11.4%	19.8%	2.4%
ABC	62%	17.5%	13.5%	7%

Table 3.11: Ranking of 'Best' Translations for NPs: chunks derived from multiple on-line MT systems

As with our sentence test set, combining fragments from multiple systems causes a deterioration in the ranking of 'best' translations for NPs. The exception to this is where chunks from systems B and C are combined. In this case we note a 6.4% improvement for the ranking of 'best' translations. Overall, the 'best' translation is consistently ranked within the top ten for NPs. For both sentences and NPs, the 'best' translation remains within the top ten for most cases and in the worst scenario, no more than the top 1% of translation candidates need to be presented to identify the 'best' translation. Accordingly, the ranking process facilitates the pruning of translation candidates output by the EBMT system.

3.4.5 Relative Gain of EBMT

Words that are not present in our system's memories cannot be translated. Web-based systems on the other hand are extremely robust and will invariably produce a translation no matter what input they are presented with. It stands to reason, therefore, that the on-line systems outperformed our EBMT system in terms of coverage. Where our EBMT system cannot translate a word, a partial translation is output with the 'missing' word inserted in the source language. While we do not consider these cases to be complete translations, this does add a level of robustness to our system.

In order to compare the quality of translations produced by our EBMT system with the quality of translations produced by the on-line systems, we translated all sentences in our test set via the individual on-line systems. We then asked two evaluators to compare the translation produced by the on-line systems and the translation produced by our EBMT system and for each sentence to state which translation they preferred. The example in (101) is from (Way and Gough, 2003) and shows some instances where the translations produced by our system were judged better than those produced by individual on-line systems. For the translations produced by our EBMT system, its memories were seeded with translation fragments derived from all three of the on-line MT systems (ABC).

(101) *Input*: Her short term interest rates link the issues.
MT A: Son lien à court terme de taux d'interêt les questions.
EBMT ABC: Ses taux d'interêt à court terme lient les questions.

Input: The researchers air the shows.
MT B: L'air de chercheurs les expositions.
EBMT ABC: Les chercheurs aérent les expositions.

Input: A group hire lawyers to provide information about clients.
MT C: Un avocats de la location du groupe fournir de l'informations au sujet de clients.
EBMT ABC: Un groupe embauche des avocats à fournir de l'informations au sujet de clients.

In all cases where both systems produced a complete translation, the translation produced by our EBMT system was preferred. In the cases where our system failed to produce a complete translation as in (89), the translation produced by the on-line system was preferred. In the first two examples in (101), we noted that our EBMT system improved the quality of the NP. Our system also produced a finite verb where the translations produced by the on-line systems had none. In the final translation in (101), we also managed to translate *hire* correctly as a verb rather than a *noun*.

We calculate the *Net Gain* of our web-based EBMT system over the on-line RBMT systems using the following formula:

(102) $$\textit{Net Gain} = \text{Coverage Percentage} + \text{K(Translation Quality)}$$

The 'Coverage Percentage' takes into account the cases where no translation is produced. The 'Translation Quality' is the number of translations preferred by the human. This figure does not include those where no translation is produced. Where we consider coverage and quality to be equally important, K=1. Under this condition, the Net Gain of EBMT is as follows:

(103) $$\begin{aligned} NetGain_{EBMT} &= 92 + 30 = 122 \text{ (compared to system A)} \\ NetGain_{EBMT} &= 92 + 8 = 100 \text{ (compared to system B)} \\ NetGain_{EBMT} &= 92 + 6 = 98 \text{ (compared to system C)} \end{aligned}$$

In the case where coverage and quality have equal priority, then our EBMT system outperforms system A *(SDL)* by a factor of 22. However, when compared against the performance of system B *(Reverso)* and system C *(Logomedia)*, our system suffers respectively no gain and a slight loss. Again, this may indicate that these systems are of a higher quality than system A.

Although our system is outperformed in terms of coverage, we have identified the instances where no translation is produced and are confident that this problem can be overcome in future models. When K=1, this indicates that quality and coverage are considered equally important. However, when we deem quality to be twice as important as coverage (i.e. K=2), we achieve an overall net gain when comparing our EBMT system to the on-line systems:

(104) $$\begin{aligned} NetGain_{EBMT} &= 92 + 60 = 152 \text{ (compared to system A)} \\ NetGain_{EBMT} &= 92 + 16 = 108 \text{ (compared to system B)} \\ NetGain_{EBMT} &= 92 + 12 = 104 \text{ (compared to system C)} \end{aligned}$$

3.4.6 Evaluating individual on-line MT systems

We used three on-line MT systems to seed the memories of our EBMT system and performed various experiments where we produced translations using various combinations of knowledge sources. Although not the primary aim of our research, as a result, we were able to evaluate the on-line MT systems used to seed our system's memories. All translations produced by our EBMT system were manually evaluated in terms of coverage and quality. We observed that combining knowledge sources did not improve coverage for our sentence test set but the number of NPs which could be translated increased from 474 to 480 when a combination of all three knowledge sources were used.

In terms of quality, a human evaluation showed that the translations produced when our system was seeded with chunks derived from *Logomedia* were approximately twice as intelligible as when *Reverso* was used — indicating that *Logomedia* (system C) is probably the best of the three on-line systems used. When combinations of chunks derived from different systems were used, we observed an increase in quality. Those combinations which integrated chunks derived from *Logomedia* performed better than those without. This pattern is true for both sentences and NPs. We also observe in section 3.4.5 that the relative gain of our system over *Logomedia* is slightly lower than for *SDL* and *Reverso*. This is further evidence that *Logomedia* is probably the better on-line system.

We noted that each of the on-line systems made consistent errors or generated correct translations for specific structures. For example, a number of the translations produced by system A incorrectly translated some adjective as verbs within the phrase. For example, in (105), the adjective *sweeping* is translated as the verb *balayer* which means *to sweep*:

(105) as is common with sweeping legislation ⇔ comme est commun avec balayer de législation

System B *(Reverso)* produced alternative translations in brackets where appropriate (cf. 106). Furthermore, where a translation was unavailable *Reverso* produced the string in English (similar to our method of inserting untranslated words).

(106) a band ⇔ une bande (un orchestre)

Logomedia produced the correct verb translation in a number of cases where the other systems failed on this measure. It is probable that this factor contributed to *Logomedia* obtaining a higher score than the other on-line systems. For example, (107a) shows the translation produced by *Logomedia* for the English phrase *by retreating to the security.*

The translation in (107b) was produced by *Reverso*. The verb *retirer* is a better translation in this context.

(107) a. en se retirant à la sécurité

b. en reculant à la sécurité

3.5 Web-based *post hoc* Validation

The WWW is a large and growing resource. There are an increasing number of Web pages appearing in multiple languages, classifying the WWW as a readily-available, potentially multilingual corpus. While not all information contained in the WWW is accurate or uses high quality language, the assumption is that the useful information will outweigh this, rendering the WWW a powerful linguistic resource.

The concept of applying the WWW to NLP stems from (Grefenstette, 1999). Grefenstette views the WWW as an extremely large corpus of attested examples and supports the idea that the size of this corpus can overcome any noise. Grefenstette believes that language models can be extracted from the WWW. Such models, he envisages, can be used to solve different NLP tasks. While other researchers (Soricut et al., 2002) ruled out querying the WWW to select and rank translations due to infeasible search time, Grefenstette is of the opinion that with increased computer memory and power, the possibility of creating useful language models from the Web is very conceivable.

Grefenstette explores the application of the WWW in choosing one translation over another. In his experiment, the entire WWW is visited using the AltaVista search engine. Competing candidates for the translations of compositional compounds are searched for and the one that is found most often is selected as the 'best' candidate. For the language direction German-English, compositional NPs were extracted from a dictionary under the following conditions:

- The dictionary entry was decomposable into two other German words found in the dictionary (compound);
- The compound term was translated in the English part of the dictionary by a two-word phrase (compositionality).

The German compositional compounds chosen for the experiment were also selected so that there would be more than one possible English translation candidate. For example, the German compound *Apfelsaft* translates as *apple juice.* Decomposing the word into two individual components, *Apfel* and *Saft* and translating these individually led to the generation of *apple juice* as a translation candidate, but also generated *apple sap* as a potential translation. By using the Alta Vista search engine to locate these conflicting translations on the WWW, *apple juice* received 13,841 hits and *apple sap* received 25 hits. Therefore, the greater number of hits occurred in the search for the correct translation. Grefenstette reports 86-87% accuracy with this experiment.

3.5.1 Determiner-Noun Agreement

In our EBMT system, when a match for a chunk cannot be found in the marker-lexicon, the chunk is subjected to a generalisation process, where its marker words are replaced by the associated tag. A search is then performed within the generalised-lexicon in an attempt to locate a match for the generalised chunk.

Where a match is found, the marker word which has been generalised is located within the word-level lexicon (assuming that it exists) and its translation is retrieved. It is then possible to insert the translation into the target side of the template. However, the marker word may have more than one possible translation, and in cases such as these only the highest weighted word will be chosen to make up the final translation. The highest weighted word is not invariably correct and it is here that problems of boundary friction may arise.

For example, assume we are translating the NP *the hard disk drives.* Given the segmentation method of the Marker Hypothesis, there are three possible ways in which this chunk may be located within our system's memories, namely those in (108):

(108) *phrasal-lexicon*: the hard disk drives
marker-lexicon: <DET> the hard disk drives
generalised-lexicon: <DET> hard disk drives

Now, assume that we have searched the phrasal and marker-lexicons and have failed to locate a match. We locate a match in the generalised-lexicon and retrieve the translation

<DET> *disques durs.* It now remains to retrieve the translation of *the* within the word-level lexicon and insert the word into the target template. Marker words such as these are likely to exist in the word-level lexicon as they occur frequently within most corpora. In French however, the translation of *the* can be one of *le, la, l'* or *les* depending on the context. As the system has no inherent linguistic knowledge to choose the correct translation at this stage, the word translation with the highest weight (in this case *la*) is chosen to produce the mistranslation in (109):

(109) *la disques durs

In this case, a feminine single determiner has been inserted into the template. Clearly the translation suffers from problem of boundary friction as the single feminine determiner *la* does not agree with the masculine plural NP *disques durs.*

In our system, we resolve this problem by integrating a *post hoc* web-based validation process. This process is based on that of (Grefenstette, 1999) described above and takes advantage of the abundance of information available on the WWW. However, while Grefenstette searches for competing candidates, our validation method is implemented by searching for the 'best translation' on-line and noting the number of hits it receives. Its morphological variants are searched for in the same way, with the assumption being that the genuinely better translation will correspond to the string which receives the greatest number of hits.

Initially, we automatically connected to search engines such as *AltaVista* and *Lycos* through the unix `'wget'` function. However, this proved to be an unreliable method as the format of the web pages was frequently changing. As an alternative measure, we opted to use the *Google*'s WEB API[12] service. This downloadable package allowed us to automatically search up to 1000 strings a day. While this limited the process to a certain extent, we found that overall it provided a reliable and consistent search engine and was sufficient for our evaluation purposes.

For each translation produced by our system, we broke down the string into trigrams and bigrams. We then searched for those sequences which could be tagged with <DET> or <POSS> and their alternatives. In the above example, we search for the sequence

[12] http://www.google.ie/apis/

la disques durs as produced by the system. However, we also search for the alternative translations *le disques durs, l'disques durs* and *les disques durs.* (110) shows the results of this search process (the number of hits each string receives using the GoogleAPI search engine).

(110) la disques durs (1 hit)
l' disques durs (2 hits)
le disques durs (37 hits)
les disques durs (39,000 hits)

Of the 500 NPs translated, 251 translations suffered from determiner-noun boundary friction. 82.5% of these (207/251) were improved *post hoc* via our Web validation method, while no alterations were made to the remaining 44 NPs. This method, therefore, can successfully identify and correct instances of determiner-noun boundary friction in the vast majority of cases for English-French. Given the abundance of information on the WWW and the growing number of multilingual documents, this approach is also generally applicable to other language pairs.

3.5.2 Noun-Verb Agreement

According to our implementation of the Marker Hypothesis as a segmentation method, verbs are not considered to be marker words. Therefore, any verbs which are encountered remain untagged and are contained within the preceding NP chunk. For example, the sentence *the boy went to the river*, would be segmented as <DET> *the boy went* <PREP> *to the river.* The verb *went* is retained with the preceding NP *the boy.*

However, given that the phrase is the largest unit to be segmented in this experiment, we do not segment sentences and therefore, the verb is not retained in context. NPs and VPs are recombined to produce a final translation for the sentences in our test set. This can lead to problems of boundary friction with regard to noun-verb agreement.

We extracted a list of all verbs in the Penn-II Treebank. We then produced translations for all verbs using the three on-line MT systems A, B and C by inserting third person singular and third person plural dummy subjects. For the French translations produced by our EBMT system, we identified the head noun as the rightmost non-marker word or the rightmost word before any other marker word in a nominal chunk. We used the list of translated verbs to identify the main verb in the sentence. We then searched for the ⟨noun,verb⟩ bigram and its morphological variant (third person plural or singular form) on the Web and corrected the translation if the alternative form received more hits than the translation produced by our system.

For example, as shown in (99), the translations for the English sentence *the boy went to the river* are:

(111) a. Le garçon allait au fleuve

b. Le garçon allaient au fleuve

c. Le garçon allait à la rivière

d. Le garçon allaient à la rivière

The translations in (111a) and (111b) both receive a score of 0.05. However, the translation in (111a) is correct as the third person singular verb form *allait* agrees with the third person singular NP *le garçon.*

Using the procedure described above, we identify *allait* and *allaient* as the main verb in (111a) and (111b) respectively. We identify *garçon* as the rightmost non-marker word in both sentences. Given this information, we search for the bigrams *garçon allait* and *garçon allaient* on the Web. The former obtains 271 hits, while the latter obtains only 7. From these results we can calculate the weights in (112):

(112)

garçon allait = (271/278)=0.975

garçon allaient= (7/278)=0.025

The probability of the correct string *garçon allait* occurring is 95% higher than the incorrect alternative *garçon allaient.* Therefore, we can use this information to raise the status of the translation in (111a) and assign it a higher ranking than the incorrect translation in (111b).

From our 200 sentence test set, we identified 58 translations which contained errors of noun-verb agreement. The results of applying our Web validation method to correct these errors *post hoc* are shown in Table 3.12.

System A: *Enterprise Translation Server*			
Improvement	No Improvement	N-V Confusion	Not found on Web
58.6%	3.4%	17.3%	20.7%
System B: *Reverso*			
Improvement	No Improvement	N-V Confusion	Not found on Web
62%	3.4%	17.3%	20.7%
System C: *Logomedia*			
Improvement	No Improvement	N-V Confusion	Not found on Web
76%	3.4%	17.2%	3.4%

Table 3.12: Using the Web to Improve Noun-Verb Agreement

Our *post hoc* Web validation process improved 34 translations for system A, 36 for system B and 44 for system C. No improvement could be made in 2 cases. We also observed that in 10 cases our methodology was unable to determine if the word to be corrected was a noun or a verb and therefore no change was made. If we were unable to locate the bigram on the web no change could be made. This problem was encountered in a small number of cases (between 2 and 12) for each system.

We can use this method to successfully validate and (in the majority of cases) improve the quality of our translations and reduce problems of determiner-noun and noun-verb boundary friction. We use the Marker Hypothesis to identify determiners and nouns and for the most part this method is successful. We can also envisage extending this process to search for and correct errors pertaining to word order, for example. Without recourse to a list of verbs in our corpus or more complex parsing techniques it may be difficult to apply the noun-verb validation process to future models. However, we predict that this particular problem will be significantly reduced when the segmentation process is applied to sentences rather than phrases. Given the increasing number of documents and multilingual information on the WWW, this validation method could certainly be extended to other language pairs. Again, depending on the languages in question, more detailed parsing may be required. Nevertheless, we deem this method to be generally applicable to solving other problems of boundary friction and portable to other language pairs.

3.6 Discussion

In this chapter we have described the development of a phrase-based EBMT system. In order to obtain a bitext, we translated 218,697 English phrases from the Penn-II treebank and translated these using three on-line MT systems. We then derived smaller aligned fragments by applying the Marker Hypothesis and generalised these fragments to facilitate the matching process. These lexical resources were used to seed the memories of our EBMT system. When the system is confronted with a new input string, it is segmented into a set of n-grams and these resources are searched. All fragments retrieved are recombined so that a set of ranked candidate translations are produced by the system. Despite using a naïve alignment algorithm and deriving the target strings in our initial bitext via on-line RBMT systems, we produce reasonable results.

We performed a number of experiments using various combinations of lexical resources to seed the memories of our EBMT system. We manually evaluated the translations produced in terms of coverage and quality, noting that when the number of fragments used to seed our system's memories increased, the performance of the system improved considerably. For our NP test set, a translation was produced for 96% of cases and 77.6% of these were correct. For 92% of cases in our sentence test set a translation was produced and 85.1% of these were deemed correct. Intelligible translations were produced in over 96% of cases for both test sets.

The translations were ranked and output with associated weights. For the majority of cases, the 'best' translation output by our system was automatically ranked within the top ten. We observed that the 'best' translation was always contained in the top 1% of translations output by the system, facilitating the retrieval of the 'best' translation by a potential user of the system.

We also compared the performance of EBMT against the three RBMT systems used to seed our system's memories. We found that in some cases, EBMT could outperform RBMT by 50%. We assessed the strengths and weaknesses of the three on-line systems and identified *Logomedia* as the 'best' system given that it outranked the others. When single MT systems were used to seed our example-base, our EBMT system performed better when *Logomedia* was used. Furthermore, when translations produced from combinations of on-line systems were examined, those generated from combinations including *Logomedia* (AC and BC) were deemed to be of higher quality.

We validated the translations produced by our system *post hoc* so that corrections in determiner-noun and noun-verb agreement could be implemented before the final translation is output to the user. We noted that this validation method was effective in improving 82.5% of translations which suffered from determiner-noun agreement and up to 76% of translations which suffered from noun-verb agreement owing to boundary friction. We have shown that despite the fact that the WWW is prone to noise, it is a sizeable resource, useful for evaluating translation candidates.

The alignment method used naïvely maps source chunks to target chunks sequentially subject to their marker tags matching. In this case, the largest unit for segmentation is the phrase and this method proves to be reasonably effective. Nevertheless, there are cases where the alignment method fails to retain potentially useful correspondences between ⟨source, target⟩ chunks. Although English and French have similar word order, there are cases where the alignment of chunks is non-sequential. This method only allows for 1:1 alignments and can, therefore, be compared to the structure-preserving translation methods of rule-based transfer systems criticised in (Hutchins and Somers, 1992). To this end it is insufficient. Further improvements to the alignment algorithm should allow it to deal with cases of 1:2, 2:1, 3:2 etc., alignments.

Storing sententially-aligned strings could also improve the performance of the system. While this may lead to more exact matches being located for sentences, applying the segmentation process to sentences could also lead to fewer problems of noun-verb boundary friction. As we do not mark verbs, they are retained in the preceding NP and therefore the translations produced are less likely to suffer from problems of noun-verb agreement. For example, when the aligned sentences in (113) are subjected to the segmentation process, the marked chunks in (114) are generated.

(113) The boy went to the river ⇔ Le garçon allait au fleuve

(114) <DET> The boy went <PREP> to the river ⇔ <DET> Le garçon allait
<PREP> au fleuve

In the above example, *went* is unmarked, as is *allait* in the associated translation. The subsequent marker words *to* and *au* come directly after the verb, leaving *The boy went* and *Le garçon allait* as a complete sub-sententially-aligned pair. Even using our current naïve algorithm the sub-sentential alignments in (115) are produced:

(115) <DET> The boy went ⇔ Le garçon allait
<PREP> to the river ⇔ au fleuve

Therefore, the chunk *the boy went* will be retrieved as a complete unit and the problem of noun-verb agreement is eliminated in this instance.

The current evaluation of translations is carried out manually. Automatic metrics generally require a set of gold standard or oracle translations against which the translations output by the system can be compared. Such a resource is not available to us given our current bitext. Although automatic evaluation metrics can be quite harsh, they will allow us to evaluate a much larger test set and therefore could benefit future marker-based models.

In chapter 4 we show how we tackle some of these issues. We also address some of the assertions of (Carl, 2003b; Schäler et al., 2003) and use controlled language specifications in our EBMT system. We apply *Logomedia* to translate a set of controlled English sentences into French. We then seed our example-base with these sententially-aligned strings and as in our phrase-based model, we apply the Marker Hypothesis to derive a set of sub-sententially-aligned chunks, generalised templates and words. In addition, we improve on the naïve sub-sentential alignment algorithm outlined in section 3.2.2. We introduce automatic metrics for evaluating our translations — we calculate BLEU scores (Papineni et al., 2002), figures for Precision and Recall (Turian et al., 2003) and word and sentence error rates.

Chapter 4

Controlled EBMT

In the previous chapter we described a phrase-based EBMT system. We seeded our example-base with English phrases from the Penn-II Treebank and their French translations derived from three on-line RBMT systems, namely *SDL*, *Reverso* and *Logomedia*. Although we obtained reasonable results, we noted that there was room for a number of improvements:

- Including sententially-aligned strings in our example-base;
- Improving our naïve sub-sentential alignment algorithm;
- Integrating automatic evaluation metrics.

In this chapter, we describe how we address the above issues. In section 4.1 we discuss controlled translation and outline our motivation for developing a controlled EBMT system. In section 4.2, we describe the novel implementation of a controlled EBMT system (Gough and Way, 2003; Way and Gough, 2004) and show how we seed our example-base with a set of controlled English sentences and their French translations.

The sub-sentential alignment algorithm described in section 3.2.2 is naïve and limited to producing alignments of a 1:1 nature. The inadequacy of this algorithm can be demonstrated by comparing its methodology with a transfer-based RBMT system where the structure of the source is imposed on the target. In section 4.2.1 we outline the developments made to our sub-sentential alignment algorithm. Following a number of alterations, we extend its scope beyond sequential 1:1 alignments. Consequently, the number of sub-sentential alignments increases and this leads to an overall improvement in translation quality.

We introduce automatic evaluation metrics and as a result can test our system on a much larger number of sentences than was previously possible. We calculate BLEU scores (Papineni et al., 2002), using the NIST MT Evaluation Toolkit[1] and Precision and Recall figures using the tools[2] outlined in (Turian et al., 2003). We also calculate word and sentence error rates (WER) and (SER) based on a standard measure of edit distance. These metrics are described in detail in section 4.3. In section 4.4, by carrying out both a manual and automatic evaluation, we assess the effects of controlling the source and target languages on translation performance. We discuss automatic evaluation in more detail and comment on these metrics in terms of the results obtained.

We also implement a number of manual alterations to our lexical resources and make slight adjustments to our system. We show how these alterations, along with improvements to our sub-sentential alignment algorithm improve translation performance. As described in section 3.5, we implement a web-based *post hoc* validation process. We rank our translations and show how the 'best' translation can invariably be located within the top ten candidates output by the EBMT system. Finally, we summarise our results and discuss what information has been gleaned from these experiments.

4.1 Controlled Translation

Controlled languages (CLs) are natural languages which are designed using restricted grammars and dictionaries. CLs can help to restrict the ambiguity and complexity associated with natural languages and can be used to aid both human and computational text processing. As such, it can be envisaged how CL applications may be useful for MT.

In recent years, the growing interest in CLs and their applications has been evidenced by a series of CLAW workshops whose theme is controlled language applications.[3] As a direct consequence of these workshops, guidelines and applications using CLs have been

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

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

[3] http://www.controlled-language.org

initiated for many languages. Several companies and organisations such as Xerox, Caterpillar and the European Association of Aerospace Manufacturers (AECMA) have developed their own versions of a simplified English. To date, however, there has been little work done in the area of controlled translation and few systems demonstrate the integration of CL techniques into the translation process.

(Carl, 2003a; Schäler et al., 2003) have claimed that controlled translation is more applicable to an EBMT environment than a rule-based one and theorise that performing controlled EBMT should yield better translation results. They point out that in a rule-based system, control must be imposed at each stage of analysis, transfer and generation before a high quality controlled translation can be produced.

Some RBMT systems, for example, Caterpillar's CTE, CMU's KANT system (Mitamura and Nyberg, 1995; Kamprath et al., 1998) and General Motors CASL and LantMark (Means and Godden, 1996) have been used to translate controlled language documentation. However, as has been noted by (van der Eijk et al., 1996), the development of such RBMT systems to produce good quality translations is complex and time-consuming, given that they are general-purpose systems attempting to derive specific, restricted applications.

As regards the area of EBMT and controlled translation, even less has been achieved. (Schäler et al., 2003) recognise the capacity for developing controlled applications using EBMT. When one considers that the quality of the translations produced by an EBMT system depends largely on the quality of the translations in the training data, it is difficult to understand why more work has not been done with CL and EBMT, as it stands to reason that if the examples are more controlled, the translations produced will be of higher quality. This, coupled with the ability of corpus-based MT to overcome the infamous 'knowledge acquisition bottleneck' (cf. section 2.1.2) and the complexities involved in developing a controlled rule-based system, make a very good case for further research in the area of controlled EBMT.

The development of controlled language applications for EBMT has been limited to a certain extent by the lack of quality controlled bitexts in existence. One cannot assume that a high quality controlled bitext can be developed by simply imposing language-specific controlled language specifications on both the source and target languages in question. For example, a common CL rule limits the length of sentences which are processed. If such a rule is to be implemented for say, ⟨German,English⟩ translation this may cause problems. It is possible that while the German source string might conform to the given word limit, compound nouns which are classified as a single 'word' in German may be translated as several words in English and this could violate the word limit restriction for the target language.

Some work has been done in the area of deriving controlled bitexts. (Hartley et al., 2001; Power et al., 2003) prompt users who are experts in a specific technical domain to build up a text within that domain. They do not need foreign language expertise. The system facilitates multiple expressions of the same underlying input in different languages, so that the resulting strings conform to a controlled language which has been specifically defined. (Bernth, 2003) replaces certain constructions within parse trees with more desirable target text, therefore constraining the output. However, this method is not suitable for our system as we do not encode such detailed structural representations.

4.2 Our Controlled EBMT System

Motivated by the claims of (Carl, 2003a; Schäler et al., 2003), we sought to develop an EBMT system in a controlled environment. We obtained a set of 1,691 English sentences from *Sun Microsystems*. These were extracted from computer manual documentation and are written according to CL guidelines.

The controlled environment enforced by *Sun Microsystems* is termed 'Sun Proof' (Akis and Sisson, 2002) and was developed with the assistance of the Institute of Applied Information Sciences (Saarbrücken). The translatability guidelines applied by Sun Proof can be divided into three categories:

- style guidelines;
- grammar rules;
- terminology.

Sun Proof contains approximately 30 style guidelines which are intended to improve the simplicity and clarity of the text submitted for translation. One guideline states that sentence length is limited to a maximum of 25 words. For example, the sentence in (117) has been rewritten from that in (116) using the Sun Proof guidelines. When the sentence in (117) was submitted for machine translation, a significant improvement was reported from when the original sentence in (116) was submitted.

(116) This chapter provides an overview of the approach to transitioning from IPv4 to IPv6 and also provides the standardized solutions to transition from IPv4 to IPv6.

(117) This chapter provides an overview of the standardized solutions that are required to make the transition from IPv4 to IPv6.

A set of grammar rules was chosen on the basis of two conditions:

- if a violation of that rule could result in a meaning shift in the original and the translation;
- if a violation could lead to a misparse in the machine translation application.

For example, subject and verb agreement is one example of a grammar error that is flagged as incorrect.

Finally, terminology which should not be used or which is less preferred is detected by Sun Proof. For example, the word *may* is flagged as 'illegal' when it is used as a replacement for *can* or *might*. When *may* was submitted for machine translation in this context it was found to consistently produce a shift in meaning.[4] For this reason, *may* can only be used in the context of granting permission.

Given the lack of controlled bitexts, we used the on-line RBMT system *Logomedia* to translate these controlled English sentences. *Logomedia* was selected as it was deemed the 'best' on-line system by our previous research (cf. section 3.4.6) and was also named by PC Magazine as the 'recommended Internet Translation Service' in April 2003.

In this way we obtained a 'controlled' bitext which we used to seed our example-base. We realise that it is more usual in a controlled translation system to control the input

[4] This was found to be the case for the target languages used in the study outlined in (Akis and Sisson, 2002). These were German, Spanish, Japanese and Chinese.

strings. As such, we are aware that our system does not strictly conform to the definitions for controlled translation as specified in (Carl, 2003a; Schäler et al., 2003). However, we believe that our approach is justified given the lack of quality controlled bitexts in existence.

For our test set, we extracted a portion of a TM also derived from *Sun Microsystems*[5] and containing text in a related domain. We performed translation from English-French and from French-English and as a result we were able to evaluate the effects of controlling the source and target language. That is, in separate experiments we controlled the stages of analysis and generation using data written according to controlled language specifications. Figure 4.1 illustrates the derivation of our training corpus and test data.

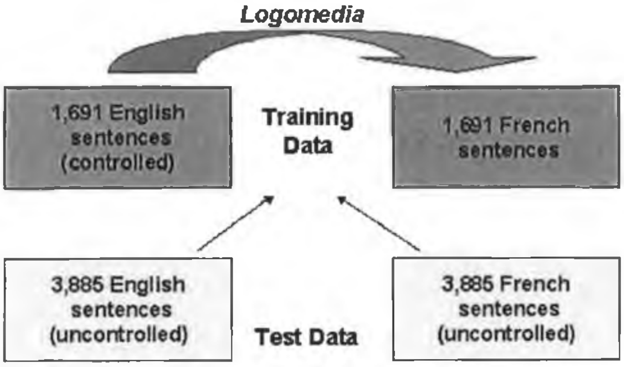

Logomedia
1,691 English sentences (controlled)
Training Data
1,691 French sentences
3,885 English sentences (uncontrolled)
Test Data
3,885 French sentences (uncontrolled)

Figure 4.1: Our Controlled EBMT system: Training and Test Data

As described in section 3.2.2 we applied the Marker Hypothesis to segment the ⟨source, target⟩ pairs and derive smaller fragments. Using the original naïve sub-sentential alignment algorithm outlined in section 3.2.2, we aligned the fragments, producing 1,079 sub-sententially-aligned chunks. We subsequently applied a novel technique where those chunks which could not be aligned using this method were translated by *Logomedia*. If the translation produced was contained in the original translation then these chunks were also aligned. For example, the segments in the ⟨source, target⟩ example in (118) cannot be aligned as the marker tags do not match and there is a different number of segments in source and target. Note that segments such as <PREP> *on* <DET> *the desktop* (under-

[5] We assume that the data in this TM was translated from English into French. However, we cannot confirm this to be the case and further research would be necessary to determine any impact of directionality on translation quality.

lined in (118)) are treated as a single segment under the caveat that each segment must contain at least one content word.

(118) <DET> an object <PREP> for <QUANT> each slice appears
<PREP> on <DET> the desktop background ⇔
<DET> un objet <PREP> pour <QUANT> chaque tranche paraît
<PREP> sur <DET> l'origine <PREP> de bureau

When we translate the individual segments via *Logomedia* we obtain the translations in (119):

(119) <DET> an object ⇔ un objet
<PREP> for each slice appears ⇔ pour chaque tranche paraît
<PREP> on the desktop background ⇒ sur l'origine de bureau

As all chunks in (119) are present in the original target sentence (118), these can be added to the marker-lexicon along with their source counterparts. Using this method of populating the marker-lexicon, we produced an additional 2082 alignments (3161 in total).

4.2.1 An Improved Sub-Sentential Alignment Algorithm

In this particular experiment, our bitext has been partially derived via *Logomedia* and therefore the method described above is a useful way of increasing the number of sub-sentential alignments in our marker-lexicon. However, the use of *Logomedia* to deduce our sub-sentential alignments is not ideal and may not work as well with an alternative bitext if one were to become available.

The alignment method applied thus far, though effective, is relatively naïve and results in a large amount of potentially useful data being discarded. This method can only be applied to segmented pairs where the number of marker tags in the source and target are the same and when these tags match sequentially. Applying these criteria, we can successfully deduce the sub-sentential alignments in (121) from the example pair in (120):

(120) <NULL> slice <DET> a partition <PREP> of <DET> a disk ⇔
<NULL>tranche <DET> une partition <PREP> d' <DET> un disque

(121) <LEX> slice ⇔ tranche

<DET> a partition ⇔ une partition

<PREP> of a disk ⇔ d'un disque

In the above example, the second chunk in the source and target sentence is marked with a <DET> tag and the final chunk is marked with a <PREP> tag in both cases. The initial word in each sentence is a non-marker word and is therefore marked with a <NULL> tag to signify this. The word correspondence derived from aligning the <NULL> chunks in the source and target is tagged with <LEX> and added to the word-level lexicon.

However, we have already observed that this algorithm is limited. When confronted with an example pair such as that in (122), further sub-sententially-aligned fragments could not be derived using the same method.

(122) <DET>$_{s1}$ the new folder resides <PREP>$_{s2}$ in <POSS> your desktop background directory ⇔

<DET>$_{t1}$ le nouveau classeur réside <PREP>$_{t2}$ dans <POSS> votre répertoire <PREP>$_{t3}$ de <DET> l'origine <PREP>$_{t4}$ de bureau

Nevertheless, it is obvious that useful information could be extracted from this example pair. Potentially, the alignments in (123) could be derived:

(123) <DET> the new folder resides ⇔ le nouveau classeur réside

<PREP> in your desktop background directory ⇔ dans votre répertoire de l'origine de bureau

That is, if source chunk *s1* (*the new folder resides*) in the English sentence could be mapped onto target chunk *t1* (*le nouveau classeur réside*) in the French sentence, then source chunk *s2* in the English sentence could be mapped onto chunks *t2*, *t3* and *t4* in the French sentence.

Given that the criteria for aligning chunks are very strict, we revised the algorithm and made a number of improvements in an attempt to retain more data and derive sub-sentential alignments from a greater number of sententially-aligned pairs. As previously, we considered that ⟨source, target⟩ chunks with the same marker tags were likely alignments. However, we also introduced a measure of lexical equivalency between ⟨source, target⟩ pairs. We translated all the words in our source language corpus via the on-line system *Logomedia* to create a base dictionary. For each ⟨source, target⟩ example pair, we used this dictionary to check for content-word equivalences between chunks. We did not check for similarity between non-content or function words because we considered that this was factored into the alignment process when matching marker tags. Those chunks in the source and target sentences which contained lexical correspondences were considered more likely alignments. An important aspect of this improved algorithm is that where previously we could only produce 1:1 alignments, we can now generate alignments of a 2:1, 3:1 etc., nature.

Table 4.1 shows the correspondences which are established between the ⟨source, target⟩ example in (122).

	S1	*S2*
T1	<new, nouveau> <folder, classeur> <resides, réside> <DET>	
T2		<directory, répertoire> <PREP>
T3		<background, origine> <PREP>
T4		<desktop, bureau> <PREP>

Table 4.1: Correspondences between source and target chunks for the example pair in (122)

Using our dictionary we can establish correspondences between *new* and *nouveau*, *folder* and *classeur*, *resides* and *réside*, *desktop* and *bureau* and *background* and *origine*. We also note that chunks *s1* and *t1* are both tagged with <DET>. All other chunks in the source and target strings are tagged with <PREP>.

We observe that chunks *s1* and *t1* share a lexical correspondence (⟨*new,nouveau*⟩) and a marker tag <DET>. Neither *s1* nor *t1* can be linked to alternative chunks. Therefore, the alignment in (124) can be established:

(124) <DET> the new folder resides ⇔ le nouveau classeur réside

Source chunk *s2* shares lexical correspondences with chunks *t2*, *t3* and *t4*. Given that these target chunks are also contiguous, they can be merged and the alignment in (125) can be produced:

(125) <PREP> in your desktop background directory ⇔ dans votre répertoire de l'origine de bureau

Therefore, where our old algorithm would fail to produce the sub-sentential alignments in (123), we can now add these chunks to our marker-lexicon.

In the above example, we derived the alignments solely from the lexical correspondences established. In other cases, it is necessary to also match the marker tags. Consider the segmented example pair in (126):

(126) <PREP>$_{s1}$ to open <DET>$_{s2}$ a drawer click <PREP>$_{s3}$ on <DET> the drawer object <PREP>$_{s4}$ in <DET> a panel ⇔
<PREP>$_{t1}$ pour ouvrir <DET>$_{t2}$ un tiroir cliquez <PREP>$_{t3}$ sur <DET> l'objet <PREP>$_{t4}$ de tiroir <PREP>$_{t5}$ dans <DET> un panneau

From this example, the word and marker tag equivalences in Table 4.2 can be derived.

	S1	***S2***	***S3***	***S4***
T1	<open, ouvrir> <PREP>			
T2		<drawer, tiroir> <click, cliquez> <DET>	<drawer, tiroir>	
T3			<object, objet> <PREP>	
T4		<drawer, tiroir>	<drawer, tiroir> <PREP>	
T5				<panel, panneau> <PREP>

Table 4.2: Correspondences between source and target chunks from the example pair in (126)

The correspondences between chunks *s1* and *t1* and *s4* and *t5* are unproblematic and the alignments in (127) can be produced. Note that we do not link chunks with common marker tags if a lexical equivalence between these two chunks has not already been established. As a result, chunks such as *s1* and *t3* cannot be linked. The naïve algorithm in section 3.2.2 aligned chunks with common marker tags without considering any lexical equivalences.

(127) <PREP> to open ⇔ pour ouvrir
<PREP> in a panel ⇔ dans un panneau

However, given that chunks *s2* and *s3* both share correspondences with chunks *t2* and *t4*, conflict arises and it is unclear as to how further sub-sentential alignments can be derived. Nevertheless, as *s2* shares a common marker tag (<DET>) with *t2* and not with *t4*, we can erase the link established between *s2* and *t4*. Similarly, we can remove the correspondence between *s3* and *t2*. Consequently, the conflict is removed and the additional alignments in (128) are produced:

(128) <PREP> on the drawer object ⇔ sur l'objet de tiroir
<DET> a drawer click ⇔ un tiroir cliquez

We also factored in the position of chunks into the alignment process. While this may not be universally applicable to all language pairs, as French and English have relatively similar word order we assumed that closer chunks were more likely to be good alignments. Consider the example in (129):

(129) <NULL>$_{s1}$ click <PREP>$_{s2}$ on <DET> the icon button <PREP>$_{s3}$ to display <DET>$_{s4}$ an icon selector dialog ⇔
<NULL>$_{t1}$ cliquez <PREP>$_{t2}$ sur <DET> le bouton <PREP>$_{t3}$ de <DET> l' icône <PREP>$_{t4}$ pour afficher <DET>$_{t5}$ un dialogue <PREP>$_{t6}$ du sélectionneur <PREP>$_{t7}$ de <DET> l' icône

Table 4.3 illustrates the lexical correspondences that can be established between the source and target chunks for this example. Based on these correspondences, we can establish the 1:1 alignments in (130).

	S1	***S2***	***S3***	***S4***
T1	<click, cliquez> <NULL>			
T2		<button, bouton> <PREP>		
T3		<icon, icône> <PREP>		<icone, icône>
T4			<display,afficher> <PREP>	
T5				<dialog, dialogue>
				<DET>
T6				<selector, sélectionner>
T7		<icone, icône> <PREP>		<icone, icône>

Table 4.3: Correspondences between source and target chunks for the sentence pair in (129)

(130) <LEX> click ⇔ cliquez

<PREP> to display ⇔ pour afficher

From the aligned bigram <PREP> *to display* ⇔ *pour afficher*, we can also derive the word-level alignment <LEX> *display* ⇔ *afficher*. However, it is unclear how additional alignments can be established given that chunks *s2* and *s4* in the source string share lexical equivalences with chunks *t3* and *t7* in the target. Chunk *s2* also shares a common marker tag (<PREP>) with both target chunks. If we consider that chunk *t3* is 'closer' to chunk *s2* than chunk *t7*, then we can eliminate the conflicting correspondences between *s2* and *t7* and *s4* and *t3*.

As a result, chunks *t2* and *t3* can be merged and *t5*, *t6* and *t7* can be merged, giving us the 1:2 and 1:3 alignments in (131):

(131) <PREP> on the icon button ⇔ sur le bouton de l'icône

<DET> an icon selector dialog ⇔ <DET> un dialogue du sélectionneur de l'icône

Of course, integrating the position of chunks in this way may also generate some incorrect alignments. Moreover, it may not be portable to other language pairs such as English-Chinese or English-Arabic. One other solution might be to produce all possible alignments given the lexical equivalences derived. In this way, although the incorrect alignment will be produced, it may not be weighted highly in our marker-lexicon.

We also consider cognates and 'close' matches to be useful. For example, lexical correspondence can be established between source and target words such as <*option,option*>, <*section, section*> and <*action,action*> when they occur in an example pair, as they are deemed cognate matches. Of course, ⟨source, target⟩ words such as <*assist,assister*> will not be matched under this condition. The example in (132) is a partial example of a segmented sentence pair in our corpus where the English sentence contains six chunks and its French equivalent, seven chunks.

(132) <PPRON> you can click <PREP> on the items... ⇔
<PPRON> vous pouvez cliquer <PREP> sur les articles...

A lexical correspondence can be established between *items* and *articles*. However, without associated context, *Logomedia* translates the verb *can* as the noun *boîte* and *click* as the imperative *cliquez*. As neither of these translations appear in our target string, it is not immediately clear how a lexical correspondence can be established between the initial chunks in the source and target examples.

However, given that the translation of *click* located in our dictionary (*cliquez*) is quite similar to the word *cliquer* in our target string (they differ by only one character), we can establish a correspondence between the source word *click* and the target word, *cliquer*.

We integrate an edit distance measure, based on *Levenshtein Distance* (Levenshtein, 1965) to measure the similarity between the words retrieved from the dictionary and the words in our target string. This measure is only implemented between words where a lexical correspondence cannot be established by the base dictionary. The Levenshtein Distance takes into account, the number of deletions, insertions and substitutions required to transform a source word (sX) into a target word (tX). The greater the distance between two strings, the more different they are deemed to be.

The *Levenshtein Distance* between *cliquer* and *cliquez* is 1. This number is compared to the average length of the strings compared. For a correspondence to be established, the formula in (133) must return a true value:

(133) $$LD \leq \frac{Ave.\ Length\ of\ string\ compared}{2} + 1$$

That is, the number returned by calculating the *Levenshtein Distance* must not be greater than half the average length of the strings plus 1. In this example, the average length of the strings compared is 7. Adding one to this figure gives us 8. At 1, the *Levenhstein Distance* is far lower than this number and therefore, the lexical correspondence <*click,cliquer*> can be formed and the initial chunks in the source and target examples can be linked.

4.2.2 Alignment Evaluation

When our original naïve alignment algorithm (cf. section 3.2.2) was applied, we produced 1,079 sub-sententially-aligned chunks from a set of 1,691 ⟨source, target⟩ sententially-aligned pairs. Using our original naïve algorithm and integrating *Logomedia* to derive translations for chunks which failed to be produced via our method, an additional 2,082 chunk pairs were produced, amounting to a total of 3,161 chunks in our marker-lexicon.

By applying the new algorithm, we populated the system's marker-lexicon with 6,400 chunks without recourse to *Logomedia.* The derivation of sub-sentential alignments previously relied on the ⟨source, target⟩ pairs meeting very strict criteria. Consequently only 18% of sentence pairs from a total of 1,691 were considered suitable candidates for generating sub-sentential alignments. When the revised algorithm is applied, we can derive sub-sententially-aligned fragments from over 87% of sentence pairs. Where sentences contained less than 2 chunks in either the English or French string, the pair was not considered useful for sub-sentential alignment. For example, the pair in (134) cannot be broken down further as the source sentence only contains a single chunk.

(134) <NULL> choose new window ⇔ <NULL> choisissez <DET> la nouvelle fenêtre

In order to assess the quality or 'correctness' of the sub-sentential alignments generated using the revised algorithm, we carried out a manual evaluation on 710 sub-sententially-aligned pairs. These were randomly extracted from a list of sub-sentential alignments produced when the algorithm was applied to the 1,691 ⟨English, French⟩ pairs in our training corpus.

We observed that when the original algorithm was applied, 586 alignments out of 710 alignments were well-formed. When the revised algorithm was applied, the number of alignments deemed to be correct fell by 7% to 536. However, because we produce more alignments in total using our revised algorithm, we observe that there is an overall improvement when coverage and quality are taken into consideration (cf. section 4.4). We also noted that all alignments which were generated correctly using our original algorithm were also produced using our revised algorithm and did not deteriorate in quality.

The quality of the translations produced by the system is dependent on the quality of the training data. Although chunk pairs which are erroneous or incorrectly aligned can improve coverage, they will ultimately have an adverse effect on the quality of the translations generated. The example in (135) shows how we can produce 2:2 alignments but also illustrates an alignment error.

(135) <PREP> to access <PPRON> your session again
<PPRON> you must enter <POSS> your password
<PREP> for <DET> this field ⇔
<PREP> pour accéder encore <PREP> à <POSS> votre session
<PPRON> vous devez entrer <POSS> votre mot <PREP> de passe
<PREP> pour <DET> ce champ

From (135), the alignments in (136) are produced:

(136) a. <PREP> to access your session again ⇔ pour accéder encore à votre session
b. <PPRON> you must enter ⇔ vous devez entrer
c. <PPRON> your password ⇔ votre mot
d. <PREP> for this field ⇔ pour ce champ

(136b) and (136d) are 1:1 alignments. (136a) is an example of a 2:2 alignment which was produced by merging the first two chunks in the source and target sentences. However, the alignment in (136c) is incorrect.

In our dictionary, the translation of password is stored as *mot de passe*. When we attempt to create a link between *password* and *mot*, we find that this is possible, based on the fact that *mot* is a complete word within the multi-word unit *mot de passe*. Therefore, a link is established between *password* and *mot*. However, given this link, no further links can be established between *password* and alternative French words, so a link between *password* and *passe* is not established.

Allowing single words to link to multi-word units such as *mot de passe* would overcome this problem. Given that each alignment is weighted (cf. section 3.4.4), we hope that better alignments will occur more often and consequently outweigh poorer alignments. It would also be useful to filter out the incorrect alignments so that they are eliminated completely from the marker-lexicon. A manual analysis of the sub-sentential alignments produced could be one way of achieving this. However, this would prove to be a tedious, labourious task for a human. An automatic filtering of the sub-sentential alignments, while potentially more error-prone, would be far more efficient. One possible solution might be to provide a length-based comparison of the alignments with a translation produced by *Logomedia*. For example, we can compare the sub-sentential alignment in (136c) with its translation via *Logomedia*. The on-line system produces the translation *votre mot de passe*. As these strings differ in length by two words, this sub-sentential alignment may be a candidate for such a length-based filtering process. However, further evaluation is necessary to determine the advantages or drawbacks of such a method. In section 5.3.3, we implement such a process and discuss its impact on the quality of translations produced by our system.

Integrating a bilingual dictionary to determine word correspondence undoubtedly helps to improve the sub-sentential alignment algorithm. However, the vocabulary provided by *Logomedia* is limited and can sometimes hinder the alignment process. For example, the English words *web* and *hide* are translated by *Logomedia* as *tissu* and *peau*. In our corpus, however, the common translations of these words are for *web*, *web* and for *hide*, *masquer* and *cacher*. Similarly, *can* appears in the dictionary as *boîte*. However, in many of the examples in our corpus, *can* is translated as a form of the verb *pouvoir*. As a result, lexical correspondences cannot be established between these words when they occur within a ⟨source, target⟩ segmented pair. Furthermore, if the system were to be extended to other language pairs, MT systems may not be available for the languages in question (e.g. English-Irish). Integrating a more domain-specific dictionary or extracting lexical correspondences from the corpus would be a means of overcoming these problems.

This may also mean scaling up our corpora to increase the reliability of such information.

4.3 Automatic Evaluation Metrics

The evaluation of translations is crucial as a means of assessing the performance of an MT system. Manual evaluation is probably the most reliable method but is a laborious, time-consuming and expensive process. Automatic evaluation metrics allow us to evaluate a far larger test set and, as such, have obvious advantages. These methods can generally be applied cheaply and efficiently and can be used repeatedly to evaluate translations and to assess the impact of additional techniques and alterations to the system. However, it is imperative that these metrics are comparable with human judgements and provide reliable and consistent results.

A number of metrics have been proposed which automate the process of MT evaluation. Some research has reported the advantages and disadvantages of these metrics (Coughlin, 2003; Turian et al., 2003). However, there are no evaluation requirements or standards for MT output and while most research now reports results using an automatic evaluation, the selection of these metrics depends solely on the developers of the system.

In order to provide as broad and extensive an evaluation as possible, we evaluated the translations output by our EBMT system using five of these metrics, namely BLEU (Papineni et al., 2002), Precision and Recall (Turian et al., 2003) and standard word and error sentence rates (WER and SER).[6] According to (Papineni et al., 2002), the closer a machine translation is to a professional human translation, the better it is deemed to be. The calculation of these metrics requires a set of gold standard or reference translations for each input sentence.[7] Essentially, the translation produced via MT is compared to this reference translation and a score is assigned to it following a comparison of the two strings. Different metrics account for different factors when measuring the similarity between a candidate and reference translation. We obtain our test set from a *Sun Microsystems* TM.

[6] We use version 09c of the BLEU evaluation software. This was downloaded from http://www.nist.gov/speech/tests/mt/resources/scoring.htm. Figures for Precision and Recall were calcualted using GTM v1.2 downloaded from http://nlp.cs.nyu.edu/GTM/.

[7] It is possible to assign more than one reference translation to each source string. The number of reference translations can affect the scores obtained in an automatic evaluation. Given the resources available, all experiments presented in this thesis were carried out with just one reference translation per source string.

As each source sentence has a corresponding translation within the TM we can use this string as a reference translation to compare against the output of our EBMT system. In this section we describe how the automatic evaluation metrics applied in our experiments assess the quality of MT output.

4.3.1 SER and WER

SER is a measure of the number of sentences in our test set which obtain an exact match with the reference translation. For example, if our test set contains 100 sentences and 10 of the translations produced are exactly the same as their corresponding gold standard or reference translation, then the overall SER for that test set is 90%. Therefore, a low SER indicates that many of the translations produced via MT obtain an exact match with a reference translation and, therefore, suggests higher quality translations.

WER is the standard evaluation metric for speech recognition systems. Its calculation is based on how much the word string returned by the system differs from the correct or reference translation. WER is calculated by computing the minimum edit distance in words between the hypothesised string (MT output) and the correct string (Reference translation) (Jurafsky and Martin, 2002):271. As a result, the minimum number of word substitutions, insertions and deletions required to map the reference translation to the translation produced by the MT system is computed:

$$(137) \qquad WordErrorRate = 100\frac{Insertions+Substitutions+Deletions}{Total\ Words\ in\ Reference\ Translation}$$

For example, consider the reference and hypothesised utterance in Table 4.3.1 from the CALLHOME corpus (Hain et al., 1998):

REF:	I	***	**	UM	the	PHONE	IS	i	LEFT	THE	portabl
HYP:	i	GOT	IT	TO	the	****	FULLEST	i	LOVE	TO	portabl
EVAL:		I	I	S		D	S		S	S	
REF:	****	PHONE	UPSTAIRS	last	night	so	the	batter	ran	out	
HYP:	FORM	OF	STORES	last	night	so	the	battery	ran	out	
EVAL:	I	S	S								

Table 4.4: Insertions, Substitutions and Deletions between a reference and hypothesised utterance from the CALLHOME corpus Hain et al. (1998)

Altogether, 6 substitutions, 3 insertions and 1 deletion can be counted for this utterance. The WER is calculates as in (138):

(138) $$WordErrorRate = 100\frac{6+3+1}{18} = 56\%$$

As is the case with SER, a lower figure for WER suggests a better result, i.e. a higher quality translation. Of course, SER and WER penalise good alternative translations which differ from the reference translations.

4.3.2 The BLEU metric

The BLEU metric is calculated based on a comparison between MT output and one or more reference translations. This comparison is based on the number of co-occurring n-gram sequences. The cornerstone of the BLEU metric is the calculation of modified n-gram precision. This score is representative of the number of n-word sequences which occur in both the reference and candidate translation. Once a matching candidate word has been identified, then the corresponding reference word is considered exhausted. Therefore, where an n-gram occurs x times in the candidate translation and y times in the reference translation and $y \leq x$ then the sequence is only counted y times. The n-gram precision p_n is calculated using the formula in (139):

(139) $$p_n - \frac{c_n \bigcap r_n}{c_n}$$

where

- c_n is the multiset of n-grams occurring in the candidate translation.
- r_n is the multiset of n-grams occurring in the reference translation.
- $|c_n|$ is the number of n-grams occurring in the candidate translation.
- $|c_n \cap r_n|$ is the number of n-grams occurring in c_n that also occur in r_n such that elements occurring j times in c_n and i times in r_n occur maximally i times in $|c_n \cap r_n|$.

Intuitively, a precision score p_n can be calculated for any value of n. (Papineni et al., 2002) consider 4 as the maximum value for n and combine scores for all values of n into a single metric. Given that longer matches are more likely to occur less frequently, as the value of n increases, the value of p_n typically decreases. In order to factor longer n-gram matches into the BLEU metric, a score p_N is calculated. This is a combined score for all values of n and is calculated by summing over the logarithm of each p_n, multiplied by weight $\frac{1}{N}$ as in (140):

(140) $p_N = \exp(\sum_{n=1}^{N} \frac{1}{N} \log(p_n))$

When p_n is calculated, any candidate translation which is longer than its reference translation is penalised. In order to penalise candidate translations which are shorter than their corresponding reference translation a *brevity penalty* (BP) is calculated and multiplied by the combined precision score p_N. A penalty of 1 is assigned where a reference translation is the same length or longer than its candidate translation. The penalty is greater than 1 when the candidate translation is shorter than the corresponding reference.

Furthermore, if candidate c_x is 1 word shorter than its reference r_x and c_y is also 1 word shorter than reference r_y but r_x is longer than r_y, then the BP for c_y should be greater than the BP for c_x. BP is calculated according to the formula in (141):

(141) $BP = e^{max(1-\frac{length(R)}{length(C)},0)}$

The BP is calculated over the entire corpus rather than calculating the BP for each sentence and finding the average. The penalty is applied to the precision score as in (142):

(142) $BLEU = BP \cdot p_N$

The BLEU score therefore, considers that a set of translations will receive a high score if the candidate translations can match the reference translation in terms of word order, similar word matches and length (Papineni et al., 2002).

4.3.3 Precision and Recall

We can define the general calculation for Precision and Recall according to the formulas in (143) and (144), assuming that C represents a candidate translation and R a reference translation.

(143) $precision(C \mid R) = \frac{C \bigcap R}{C}$

	C	B	A	I	C	D	E
A			●				
B		●					
C	●				●		
D						●	
E							●
F							
B		●					
A			●				
I				●			
C	●				●		

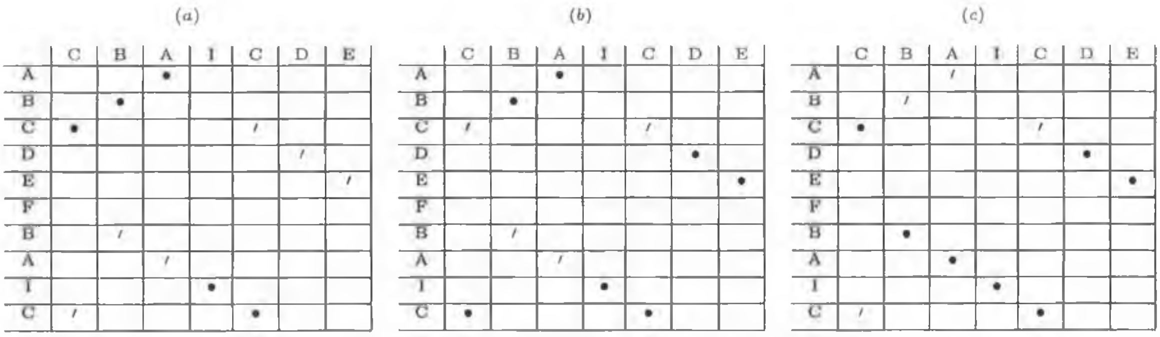
Figure 4.2: Bitext grid adapted from (Melamed et al., 2003) which shows points of intersection between a candidate translation and its associated reference translation

(144) $$recall(C \mid R) = \frac{C \bigcap R}{R}$$

By defining a method of calculating the intersection between a candidate and a reference translation (Melamed et al., 2003; Turian et al., 2003) apply these metrics to evaluate MT output. Figure 4.2 is adapted from (Melamed et al., 2003; Turian et al., 2003) and illustrates the intersection of two texts. The top of the grid represents the candidate translation from left to right. The reference translation can be read at the left-hand side of the grid from top to bottom. The *hits* or points of intersection between the candidate and reference strings are indicated by bullet points in the grid.

The number of bullet points which appear in a column is consistent with the number of *hits* for the candidate word at the top of that column. We can see from Figure 4.2 that the candidate word C represented at the left-most column obtains two hits as the word C occurs twice in the corresponding reference translation. According to this pattern, the calculation of (C $\bigcap$ R) is an over-estimation. In order to overcome this problem and ensure that the intersection count is not misrepresented, (Melamed et al., 2003; Turian et al., 2003) introduce the concept of *matching* so that no more than one hit appears in each row and column. Where all possible candidate words obtain a hit, this is referred to as a *maximum matching.* The number of hits in a maximum matching is referred to as the *maximum matching size* (MMS) and cannot exceed the length of the shorter string, whether this is the candidate or reference sentence.

In Figure 4.3, *(a)* is not a maximum matching. The MMS for (*b*) and (*c*) is 7.



(*a*)

	C	B	A	I	C	D	E
A			●				
B		●					
C	●				/		
D						/	
E							/
F							
B		/					
A			/				
I				●			
C	/				●		

(*b*)

	C	B	A	I	C	D	E
A			●				
B		●					
C	/				/		
D						●	
E							●
F							
B		/					
A			/				
I				●			
C	●				●		

(*c*)

	C	B	A	I	C	D	E
A			/				
B		/					
C	●				/		
D						●	
E							●
F							
B		●					
A			●				
I				●			
C	/				●		

Figure 4.3: (a), (b) and (c) are examples of *matchings* for the grid in Figure 4.2. Hits which were in the original grid but are not contained in the matching are marked /. In each matching, each row and column in the grid contains a single hit. (This illustration is adapted from Figure 1 of (Melamed et al., 2003)).

When MMS defines the intersection between a candidate and a reference translation, Precision and Recall can be calculated according to the formulae in (145) and (146) respectively.

(145) $precision(C \mid R) = \frac{MMS(C,R)}{C}$

(146) $recall(C \mid R) = \frac{MMS(C,R)}{R}$

When we consider that Figure 4.3(c) illustrates a contiguous match for four words compared to a two word contiguous match in Figure 4.3(b) it does not seem fair for both to receive similar precision and recall scores. However, given that both Figure 4.3(b) and Figure 4.3(c) obtain an equal number of hits, according to the formulae in (145) and (146) (b) is not rewarded for correct word order.

In order to account for this, (Melamed et al., 2003; Turian et al., 2003) treat runs of contiguous words as atomic units. For each such run, a block is formed and this represents the minimum enclosing square for that run. For example, the runs in Figure 4.3(b) and (c) can be converted to the blocks of cells illustrated with circles in Figure 4.4(b) and (c).

MMS can now be calculated according to the formula in (147). The aligned block area is now used to calculate the intersection between the reference sentence and the candidate translation. A single run is defined as the square of its length.

(b)

	C	B	A	I	C	D	E
A		○	○				
B		○	○				
C							
D						○	○
E						○	○
F							
B							
A							
I				○	○		
C	○			○	○		

(c)

	C	B	A	I	C	D	E
A							
B							
C	○						
D						○	○
E						○	○
F							
B		○	○	○	○		
A		○	○	○	○		
I		○	○	○	○		
C		○	○	○	○		

Figure 4.4: (b) and (c) are examples of maximum matchings for the grid in Figure 4.2. (This illustration is adapted from Figure 1 of (Melamed et al., 2003).)

(147) $$MMS(M) = \sqrt{\sum_{r \in M} length(r)^2}$$

As Precision and Recall are calculated according to the formulae in (145) and (146), the translation represented in (*c*) is rewarded and assigned a higher score to that in (*b*).

4.4 Experiments and Results

We performed a number of experiments to test our controlled EBMT system. For English and French, we randomly extracted 3,885 sentences from an uncontrolled *Sun Microsystems* TM as a test set. We ensured that all the words contained in these sentences existed somewhere in the word-level lexicon created via the Marker Hypothesis. Any words that were not located within the word-level lexicon were translated via *Logomedia* and added to the lexical resource. For example, the word *panel* appears in our test set but not in our word-level lexicon. We apply *Logomedia* to derive a translation *panneau* and the word-level alignment ⟨*panel, panneau*⟩ is subsequently added to the word-level lexicon. This ensured that for each sentence in the test set, we could at least output one translation, albeit word-for-word.

In an initial experiment, we performed controlled generation and controlled analysis by translating our entire test set from French-English and English-French. As we pointed out in section 4.2, we are aware that this approach deviates from the norm but we suggest that it can be justified given that no suitable controlled bitext was available to us. We carried out a detailed automatic and manual evaluation on the translations obtained.

When deriving our lexical resources via the Marker Hypothesis, we initially applied the naïve alignment algorithm described in section 3.2.2 and then our new improved algorithm (cf. section 4.2.1). We saw in section 4.2.2 that the revised alignment algorithm had a positive effect on the coverage of the alignments produced and an overall improvement was noted in alignment quality. By seeding our system's memories with the sub-sentential alignments derived via the revised algorithm we will show that translation performance also improves.

We integrated a number of amendments in an attempt to improve the performance of our system. Firstly, we made some corrections to our word-level lexicon and secondly, we allowed for multiple word translations to be used in generating the translations for each sentence. We will show how these minor alterations improve the quality of translations output by the EBMT system.

We also translated our entire test set directly via *Logomedia* so that we could compare the results of EBMT to those of an RBMT system. We performed both an automatic and manual evaluation of the translations produced by our system and by *Logomedia.* Our objective is to substantiate the claims of (Carl, 2003a; Schäler et al., 2003) that EBMT is more suited to controlled translation than RBMT.

In summary, we performed the following experiments and at each stage we compared our results with *Logomedia*:

- **French-English: *Controlled Generation***
  Controlling the Target Language (Alignment 1)
  Controlling the Target Language (Alignment 2)
  Controlling the Target Language (Novel Improvements)

- **English-French: *Controlled Analysis***
  Controlling the Source Language (Alignment 1)
  Controlling the Source Language (Alignment 2)
  Controlling the Source Language (Novel Improvements)

Finally, we discuss the results obtained and perform an analysis of the automatic evaluation metrics.

4.4.1 Controlling the Target Language: French-English

Automatic Evaluation

For the translations produced from our 3,885 sentence test set, we calculated BLEU scores, Precision and Recall figures and WER/SER (cf. section 4.3). The results obtained when the original naïve sub-sentential alignment algorithm was used to seed our system's memories is in Table 4.5.

Experiment	*Precision*	*Recall*	*BLEU Score*	*WER*	*SER*
Alignment 1	0.1815	0.3183	0.0836	96.7	98
Logomedia	0.2617	0.3601	0.1637	98.1	96

Table 4.5: Summary of results for controlling the target language (French-English) for the original naïve alignment algorithm using Automatic Evaluation Metrics

Using these automatic evaluation metrics we found that our EBMT system was outperformed considerably by *Logomedia.* Although we received a slightly better WER (96.7 compared to 98.1), the average BLEU score for our system was 0.0836 compared to 0.1637 for *Logomedia. Logomedia* also outperformed our system in terms of Precision and Recall.

When we applied our new improved sub-sentential alignment algorithm (cf. section 4.2.1) we noted a 44% improvement in the BLEU score for our system. Table 4.6 shows these results.

Experiment	*Precision*	*Recall*	*BLEU Score*	*WER*	*SER*
Alignment 2	0.2641	0.3211	0.1204	88.7	96
Logomedia	0.2617	0.3601	0.1637	98.1	96

Table 4.6: Summary of results for controlling the target language (French-English) for revised alignment algorithm using Automatic Evaluation Metrics

Previously, the best BLEU score for a single sentence was 0.9131. This score now rises to 1.000. Precision and Recall also improve, as do WER and SER. We now outperform *Logomedia* (although not significantly) in terms of Precision and our figures for WER and SER are also better. However, *Logomedia* continues to outperform our system in terms of Recall and BLEU score.

We made a number of amendments to our lexicons in an attempt to increase the

BLEU score for our system. We identified words in our test set that occurred more than 10 times (cf. Table 4.7). This amounted to 10% of total words. We then corrected any of these words which had been mistranslated (64 words in total) and re-ran our translations. Following this minor amendment, we noted that the BLEU score increased to 0.1267. Encouraged by this increase of 5% from our baseline score of 0.1204, we corrected those erroneous words which occurred more than once in the test corpus (30% of total words). We noted a further 20% increase in the BLEU score for our system. However, even at 0.1449, it is still outperformed by *Logomedia*.

Word	*No. of Occurences*	*Mistranslation*	*Amended Translation*
hide	25	peau	masque
web	10	tissu	web
password	5	mot	mot de passe

Table 4.7: Examples of words in our lexicon which were amended

In a final amendment to the system, we reviewed the generation of the translations. When translating sentences, a translation is located for each word in the sentence. However, it is sometimes the case that a word has multiple possible translations. The 'best' or most highly-weighted of these is chosen (cf. section 3.4.4). However, if the weighting is the same for numerous translations then one will be chosen at random.

Given that our corpus suffers from data-sparseness, the 'best' word translations are not always used in producing the final translation for a sentence. For example, the translation of the word *file* by *Logomedia* is *dossier*. However, in our set of reference translations, *fichier* is the correct translation for *file* in the majority of cases. Our word-level lexicon does contain the word correspondence <*file, fichier*> but this translation pair occurs less frequently than <*file, dossier*> and as a result it is not selected as the highest-scoring translation. To this end, we adjusted the algorithm so that the top five most frequently occurring word translations were considered as candidates for each source word. As a result, we noted an improvement of 41.7% over the baseline BLEU score. A summary of results for the automatic evaluation of French-English translations is shown in Table 4.8.

Experiment	*Precision*	*Recall*	*BLEU Score*	*WER*	*SER*
Alignment 1	0.1815	0.3183	0.0836	96.7	98
Alignment 2	0.2641	0.3211	0.1204	88.7	96
top ten% 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	98.1	96

Table 4.8: Summary of results for controlling the target language (French-English) using Automatic Evaluation Metrics

Manual Evaluation

We also carried out a manual evaluation on 200 translations which were randomly extracted from our 3,885 sentence test set. We evaluated the translations produced when our naïve alignment algorithm was used. Following the application of the new sub-sentential alignment algorithm and the novel improvements made to our system, we performed a new evaluation to confirm the positive impact of our novel adjustments on translation quality and to ensure that our results corresponded to those derived using the automatic evaluation metrics.

The manual evaluation metrics which determined the quality of a translation in our phrase-based model were determined intuitively and were defined by a single measure (quality). The metrics described in this section differ from the original scale (cf. p.70), in that the overall quality of a translation is determined in terms of both intelligibility and accuracy. Intelligibility accounts for any grammatical errors, mistranslated words etc., while accuracy ensures that the translation produced by the system is in fact a true reflection of the content of the input string.

We measure intelligibility using a 4-point scale:

- Score 3: Very Intelligible (intelligible translation, no syntactic errors);
- Score 2: Adequately Intelligible (intelligible translation, minor syntactic errors);
- Score 1: Only Slightly Intelligible (poor translation, major syntactic errors);
- Score 0: Unintelligible.

We measure accuracy on a 5-point scale:

- Score 4: Very Accurate (good translation, represents source faithfully);
- Score 3: Quite Accurate (accurate translation, minor errors of fidelity);
- Score 2: Reasonable Accurate (accurate translation, average no. of errors of fidelity);
- Score 1: Barely Accurate (poor translation, major errors of fidelity);
- Score 0: Inaccurate Translation.

The review of the manual evaluation metrics came about following a study of the reliability of such metrics in (Dabbadie et al., 2002). Although we did not wish to deviate dramatically from the original scale, we aimed to improve the reliability and coherence of the metrics used by combining our intuitive measures with those suggested in (Dabbadie et al., 2002).

System	*Score 0*	*Score 1*	*Score 2*	*Score 3*	*Exact Match*
Alignment 1	10	30	35	118	7
Alignment 2	4	12	46	126	12
Logomedia	2	21	40	123	14

Table 4.9: Comparing our EBMT system with *Logomedia* when controlling the target language (French-English) using Manual Evaluation Metrics: Intelligibility

The results for intelligibility are presented in Table 4.9. When using our naïve alignment algorithm, *Logomedia* outperforms our system in terms of intelligibility (2.5% more score 3 translations). However, when we apply the revised algorithm and integrate novel improvements to our system, we obtain 1.5% more translations with a score 3 than *Logomedia*. Only 2% of the translations are considered unintelligible. With our old algorithm this figure was 5%. When we consider all translations which are adequately or very intelligible, i.e. those with a score 2 or 3, our EBMT system obtains 184 (92%) such translations. This compares favourably with *Logomedia* which obtains just 177 (88.5%) translations with a score 2 or 3. Therefore, improving the sub-sentential alignment algorithm and adjusting our lexical resources brings about an improvement in translation performance.

The results for accuracy are given in Table 4.10. Although our system only obtains 12 exact matches in comparison with 14 for *Logomedia*, we outperform the on-line system

System	*Score 0*	*Score 1*	*Score 2*	*Score 3*	*Score 4*	*Exact Match*
Alignment 1	9	30	19	42	93	7
Alignment 2	2	6	18	36	126	12
Logomedia	9	27	27	31	92	14

Table 4.10: Comparing our EBMT system with *Logomedia* when controlling the target language (French-English) using Manual Evaluation Metrics: Accuracy

and note a significant improvement in the results when our new alignment algorithm is applied and adjustments are made to our system. When the naïve alignment algorithm is used, 71% of our translations receive a score 3 or 4, compared to 68.5% for *Logomedia*. When our new algorithm is implemented, the number of translations obtaining a score of 3 or 4 rises to 87%.

Summary of Results for Controlling the Target Language: French-English

We have shown that by improving our sub-sentential alignment algorithm, amending our lexical resources and adapting the algorithm to consider multiple translations for words, our EBMT system can outperform the rule-based system *Logomedia*. We obtain a BLEU score 0.66% higher than the rule-based system and our scores for Precision and Recall are better than *Logomedia* by about 4% and 0.45% respectively. Our results improve incrementally as a new sub-sentential alignment algorithm is implemented and adjustments are made to our system. The manual evaluation carried out to test intelligibility and accuracy supports our automatic evaluation and indicates that these are harsh measures in assessing translation quality. We provide further discussion of the relative merits of the automatic evaluation metrics in section 4.4.4. We outperform *Logomedia* by 3.5% in terms of intelligibility and by 18.5% in terms of accuracy.

4.4.2 Controlling the Source Language: English-French

Automatic Evaluation

Using the same techniques — implementing a revised sub-sentential alignment algorithm and adjusting our lexical resources — we controlled the source language by performing translation from English-French. The results are presented in Table 4.11.

Experiment	*Precision*	*Recall*	*BLEU Score*	*WER*	*SER*
Alignment 1	0.3081	0.4477	0.0925	71.8	93
Alignment 2	0.3115	0.4566	0.0954	70.0	92
top ten% 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

Table 4.11: Summary of results for controlling the source language (English-French) using Automatic Evaluation Metrics

Despite the incremental improvements to our system, *Logomedia* continues to outperform our system in terms of BLEU score (0.2321 compared to 0.1352). However, we outperform the on-line system in terms of Precision and Recall. Moreover, although BLEU suggests that translating into English produces superior results, according to precision, recall, WER and SER, performing translation in this direction (English-French) is more successful. This raises a number of interesting questions; Objectively, translating from French-English should be less problematic than translating from English-French. The French language presents more instances of agreement between, for example, determiners and nouns. In this case however, translating from English-French yields higher quality translations, at least according to the majority of automatic evaluation metrics. This in turn highlights another point of interest — the apparent anomaly between the different metrics used in our automatic evaluation. Following a manual evaluation, in subsequent sections we examine these issues in more detail.

Manual Evaluation

Again we carried out a manual evaluation of 200 sentences using the same scales of intelligibility and accuracy as in section 4.4.1. We find that *Logomedia* outperforms our system in terms of intelligibility. We obtain 188 (94%) translations with a score 2,3 or exact match. For *Logomedia*, this figure is 195 (97.5%). We outperform *Logomedia* in terms of accuracy. 90% of our translations receive a score of 3 or higher. For *Logomedia*, only 80% of translations receive such a score. The results of our manual evaluation support those of Precision, Recall and WER/SER and consequently, provide further evidence in favour of controlling the source language and performing English-French translation.

4.4.3 Controlling the Source Language versus Controlling the Target Language

When we compare the results obtained for French-English and English-French, we see that in terms of BLEU score, our system is approximately 26% less successful when translating from English-French than for French-English. *Logomedia* also outperforms our system for this language direction. We also observe that *Logomedia* obtains a BLEU score approximately 4.2% higher when translating from English-French than it does for French-English.

However, we can see from these results that the BLEU scores and the Precision and Recall figures do not corroborate. While BLEU suggests a better performance from French-English, in terms of Precision and Recall our system significantly improves in the direction English-French. When translating from English-French, Precision and Recall return percentage values of 39% and 53% respectively. From French-English, these figures are lower at 30% and 36%. We also outperform *Logomedia* from English-French in terms of Precision and Recall (Precision 35.5%, Recall 37%).

WER and SER corroborate the figures for Precision and Recall and a manual evaluation also endorses these results. This would suggest that the BLEU scores are anomalous. Taking the figures for Precision and Recall into consideration and excluding the BLEU scores, we observe that we outperform *Logomedia.* These figures also indicate that controlling the source language (English-French) produces better results using our EBMT system. This is surprising, given that translating from English-French is more likely to raise problems of boundary friction and ultimately reduce the quality of the translations produced. For example, the English determiner *the* has four alternative translations in French — *le*, *la*, *l'* and *les.* The highest weighted translation which is inserted into the target side of the template has a one in four chance of being correct in that context. Furthermore, it is very possible that the final highest weighted translation (which is submitted for automatic evaluation) will contain errors due to boundary friction, word order etc. This in turn will have an negative impact on the figures produced from an automatic evaluation. In contrast, when performing translation from French-English, only one word *the* is a candidate translation for all four of the French words, *le*, *la*, *l'* and *les.* Consequently, in this case, the problem of determiner-noun boundary friction does not occur in the English translation.

In the following section, we provide a more detailed analysis of the automatic evaluation metrics, with particular emphasis on the BLEU scores obtained.

4.4.4 Discussion of Evaluation Metrics

Calculating automatic evaluation metrics such as BLEU or Precision and Recall requires a set of reference translations which provide a prototype for a 'good' translation. Given that we extracted our test set from a *Sun Microsystems* TM, we were able to identify an oracle reference translation for each test sentence.

Previously when automatic evaluation metrics were unavailable to us, we carried out a manual evaluation on just 200 sentences and 500 NPs. Integrating automatic evaluation allows us to evaluate a far larger test set (3,885 sentences) and, as such, it is a valuable tool, the benefits of which cannot be overlooked.

As the main reason for using automatic evaluation is to reduce the amount of time a human spends evaluating translations, automatic evaluation metrics should correlate with human judgement. However, if there is a large disparity between a human evaluation and the figures provided by automatic metrics such as BLEU and Precision and Recall, then automatic evaluation would become less useful. Moreover, if the figures obtained by the various automatic evaluation metrics do not coincide then how are we to know which one to apply? Recently, there has been a great deal of focus on the use of automatic evaluation metrics in MT. (Turian et al., 2003) discussed the merits of BLEU, NIST and f-measures in MT evaluation and found the latter to be the most reliable metric. (Coughlin, 2003) found BLEU and NIST to be more reliable but emphasised the underlying importance of human evaluation. In addition, some alternative methods for automatically evaluating translations have been proposed. (Kulesza and Shieber, 2004) suggest a new class of metrics which use machine learning techniques to evaluate translations. In some initial experiments they report an improvement on existing metrics and an increased correlation with human judgement.

In the experiments discussed in section 4.4, we find that there is a disparity between the figures obtained for BLEU and the alternative evaluation measures. Precision and

Recall figures show that English-French translation yields better results. WER and SER support this hypothesis and a human evaluation of 200 sentences measuring Accuracy and Intelligibility also corroborate these figures. On the other hand, the BLEU scores indicate that performing translation in the direction French-English generates higher quality translations. We note, therefore, that our results substantiate the findings of (Turian et al., 2003) who report Precision and Recall measures to be more reliable in evaluating translation quality than BLEU.

We propose several reasons for the irregularity in the BLEU scores. Instances of agreement in gender and number in French are more common than in English. Translating into French also gives rise to verb forms which are more morphologically rich than their English counterparts. Although we correct some determiner-noun agreements via our *post hoc* validation method (cf. section 3.5), there remain some agreement errors in translations submitted for evaluation. While such translations may appear intelligible, BLEU can often return a score of 0 and in this sense, we find it to be an unduly harsh metric. Consider the example in (148). It shows the translation produced by our system for an English sentence and includes the oracle or reference translation for comparison.

(148) *Source*: those used to locate network users
Translation: ceux utilisées pour localiser réseau utilisateurs
Reference: ceux utilisés pour localiser les utilisateurs du réseau

Although the translation output by the EBMT system is intelligible, it contains a number of errors:

- incorrect agreement between the masculine plural determiner *ceux* and the feminine plural past participle *utilisées*;
- incorrect translation of *network users* as *réseau utilisateurs*.

Given these errors, we would not expect the BLEU score to be very high for the translation in (148). On the other hand, we would not expect the same translation to receive a BLEU score of 0. However, using the NIST evaluation toolkit, we find that a BLEU score of 0 is returned when this translation is submitted for evaluation. This measure would appear to be relatively harsh. When we evaluate the same translations in terms of WER, we obtain a figure of 50%. In terms of Precision and Recall, the results returned are 0.6250 and 0.8330 respectively. That is, when we compare the words in our output translation to those in the reference translation, we note that we find five of them (i.e. $\frac{5}{8}$). Of the six words in our output translation, five of these appear in the reference translation (i.e. $\frac{5}{6}$). The BLEU score attributed to this translation, therefore, appears unduly severe. Now consider the French-English translation in (149):

(149) *Source*: ceux utilisées pour localiser les utilisateurs du réseau
Translation: those used to locate the mouse pointer users to network
Reference: those used to locate the network users

The correct translation of *ceux utilisées* as *those used* is produced, given that agreement is less problematic when translating from French-English. However, the output translation contains the following errors:

- *les utilisateurs* is translated as *mouse pointer users* due to incorrect alignments in the marker-lexicon;
- the word order in the output translation is incorrect.

As a result of these errors, the English translation is less intelligible than the French translation produced in (148). Surprisingly, however, where the translation in (148) obtained a score of 0, the translation in (149) obtains a BLEU score of 0.2907. This is mostly due to the n-gram *those used to locate* which is present in both the translation output by the system and the reference translation.

It would also appear that making minor adaptations to the output translation can result in a considerable change in the BLEU score. For example, consider the French translation in (150)

(150) *Source*: execute a command on another machine
Translation: exécutez un ordre sur un autre machine
Reference: exécuter une commande sur une autre machine

This is an intelligible translation. The bigram <*un, autre*> contains an agreement error, as *un* is a masculine determiner in conflict with a feminine noun, *machine.* Again, the BLEU score of 0 which it obtains appears unduly harsh. Moreover, the figures for Precision and Recall (0.4285 and 0.4285 respectively, i.e. $\frac{3}{7}$) and WER (57%) indicate that this is far from an unacceptable translation and would appear to contradict the BLEU score.

We made a number of amendments to the translation output by the system and resubmitted it for evaluation to assess the impact that this would have on the automatic evaluation metrics. Firstly, we amended the bigram <*un, autre*> to <*une autre*> as in (151):

(151) *Source*: execute a command on another machine
Translation: exécutez un ordre sur une autre machine
Reference: exécuter une commande sur une autre machine

As illustrated in Table 4.12, we observe that on evaluation of the amended translation, the BLEU score increases from 0 to 0.4111. Although the WER falls, the figures for Precision and Recall also increase.

WER	42.8%
SER	100%
BLEU Score	0.4111
Precision	0.5714
Recall	0.5714

Table 4.12: Automatic Evaluation scores obtained for the data in (151)

The translation of *command* produced by the EBMT system is *ordre.* This is perfectly intelligible and well-formed. However, the translation of *command* present in the reference string is *commande.* When we alter the translation output to match the reference translation as in (152), we observe that the BLEU score now increases to 0.8091. Table 4.13 provides a complete list of the results obtained.

(152) *Source*: execute a command on another machine
Translation: exécutez une commande sur une autre machine
Reference: exécuter une commande sur une autre machine

WER	14.3%
SER	100%
BLEU Score	0.8091
Precision	0.8571
Recall	0.8571

Table 4.13: Automatic Evaluation scores obtained for the data in (152)

There are a number of additional points to note in relation to the automatic evaluation metrics.

- The improved performance of *Logomedia* from English-French can possibly be attributed in part to the fact that it is a rule-based system and therefore is less likely to suffer from problems of boundary friction;

- The word-level lexicon was partly derived via *Logomedia.* Any words within our test set which were not added to the word-level lexicon via the Marker Hypothesis method were translated using *Logomedia* and added to the word-level lexicon. As the controlled English has more words in common with the test set than the uncontrolled French, a larger quantity of words need to be translated in this way from French-English than for English-French. This also contributes to the improved performance from English-French;

- In general, matching from English-French is likely to outperform matching from French-English given that 'good' quality data is being compared from English-English. Matching from French-English on the other hand involves comparing good quality uncontrolled data against data derived via *Logomedia*;

- Although the controlled training set and the uncontrolled test set are of a similar domain, the amount of overlap is considerably reduced given that our test set is far larger than our training data. An increased example-base should increase the probability of locating similar chunk matches and consequently increase the BLEU score.

4.5 *Post Hoc* Validation

We use the same method as described in section 3.5 to identify determiner-noun bigrams and correct instances of boundary friction by searching for alternative candidates on the Web. For example, (153) shows a translation produced by our system for the English sentence *the network is from the list*:

(153) la réseau est de la liste

This string was produced when a generalised template in (154) was retrieved from the system's generalised-lexicon.

(154) <DET> network is from the list ⇔ <DET> réseau est de la liste

The translation of *the* which has the highest weighting in our word-level lexicon is the feminine singular determiner *la*. This was inserted into the target template to produce the translation in (153). However, we can see that this translation suffers from boundary friction as the feminine singular determiner *la* clashes with the masculine singular noun *réseau*.

We can identify the alternative bigrams in Table 4.14 and as described in section 3.5, we use *Google's WEB API service* to search for these strings. Although the bigram *la liste* does not suffer from boundary friction, it is selected as a candidate for validation and its alternatives are also searched for on the Web.

Bigram	***No. of Web Occurrences***
la réseau	84
le réseau	39,100
les réseau	56
l'réseau	24
la liste	35,000
le liste	1340
l'liste	772
les liste	149

Table 4.14: No. of Web Occurrences for determiner-noun bigrams

The correct forms, *le réseau* and *la liste* receive 39,100 and 35,000 hits respectively — far higher than the alternative erroneous candidates. The bigrams are corrected and validated and the translation in (155) is output by the system *post hoc*.

(155) le réseau est de la liste

In an evaluation of the *post hoc* validation method, we submitted 1,588 French determiner-noun bigrams for validation. 1,512 of these strings were deemed correct. For the remaining 76 cases, 53 (69.7%) were corrected. The remaining 23 strings were unchanged.

In our phrase-based system, we performed a measure of noun-verb validation (cf. section 3.5.2). This was facilitated by a list of verbs contained in the Penn-II Treebank and their translations derived via *Logomedia*. However, where previously the phrase was the longest unit stored in our example-base, we now store sententially-aligned pairs. For example, in our phrase-based system, a sub-sentential alignment such as that in (156) could potentially exist in our marker-lexicon.

(156) is from the list ⇔ est de la liste

Applying the Marker Hypothesis to our sententially-aligned strings, however, ensures that it is not possible for such an alignment to be generated. For example, the string *le réseau est de la liste*, would be segmented as in (157):

(157) <DET> le réseau est <PREP> de la liste

Given that verbs are not classified as marker words, *est* is unmarked, and *le réseau est* is retained as a complete unit. Therefore, the correct agreement between the singular masculine noun *réseau* and the singular masculine determiner *le* remains intact within the sub-sentential string. The corresponding English translation, *the network is from the list* would be segmented in the same manner and *the network is* would be retained as a complete unit. When the sub-sentential alignment algorithm is applied, a link can be established between *network* and *réseau* and the chunks can be aligned.

4.6 Ranking Translations

As described in section 3.4.4, we rank the translations output by our system. We analysed the weights of the translations produced using our revised sub-sentential alignment algorithm to assess where the 'best' translation was ranked by the system. In our phrase-based system a maximum of 2000 translation candidates were generated for a single sentence. In our controlled EBMT experiment, we implemented a novel method where the low-ranking translations are pruned on the fly. This means that the highest number of translations produced for a sentence is 123. The average sentence takes approximately 0.84 CPU seconds to process. The results for both language directions appear in Table 4.15.

	French-English	**English-French**
Ranked 1	88%	85.5%
Ranked 2-10	11%	13.5%
Ranked > 10	1%	1%

Table 4.15: Relative ranking for translations produced using our new improved alignment algorithm

In the majority (99%) of cases for English-French, the 'best' translation is ranked within the top ten translations output by our system. The 'best' translation is ranked first by our system in 88% of cases for the same language direction. There is a similar pattern for French-English translations. Where the 'best' translation is ranked first by our system 85.5% of the time, it normally occurs within the top ten translations.

When we increase the number of words in our word-level lexicon, there is a decrease in the number of 'best' translations ranked first by the system. Table 4.16 illustrates these results.

	French-English	***English-French***
Ranked 1	80.2%	75.5%
Ranked 2-10	14.8%	15.5%
Ranked > 10	5%	9%

Table 4.16: Relative ranking for translations produced using our new improved alignment algorithm and additional word translations

This result is to be expected as using more words provides the system with more options and therefore leads to more translations. The maximum number of translations produced is now 200 for a sentence. The best translation occurs as the highest weighted translation for 80.2% of cases when translating from French-English. This figure is lower for English-French as there are more options for different verb forms, singular plural-nouns etc. in the French language.

4.7 Discussion

In this chapter, we have presented the first controlled EBMT system. An on-line RBMT system is used to derive French translations for a set of English strings written according to controlled language specifications. In a novel experiment, we filter the derivation of the target string using data written according to controlled language specifications.

By improving our sub-sentential alignment algorithm and making some minor adjustments to our lexical resources, we show that our EBMT system can outperform a good on-line system, *Logomedia*. We consider our results to be encouraging. Despite the fact that our system does not conform to the definition of controlled translation as outlined by (Carl, 2003a; Schäler et al., 2003), and although *Logomedia*, our baseline comparison, is not trained on our data set, we are confident that our results support the hypothesis that EBMT systems should outperform rule-based systems in a controlled environment. Furthermore, if a suitable bitext were to become available, our results show that there is scope for more extensive research in the area of controlled EBMT.

We carry out both an automatic and a manual evaluation on the translations obtained. The integration of automatic evaluation metrics allows us to evaluate a far larger test set than was previously possible. The results for Precision and Recall and WER/SER suggest that the EBMT system produces higher quality translations when the source language is controlled, i.e. when translating from English-French. Although the BLEU scores indicate that controlling the target language is better, this is in conflict with all other automatic evaluation measures and does not correlate with our manual evaluation. Further analysis of the automatic evaluation metrics leads us to suggest that the BLEU scores are anomalous. We conclude, therefore, that performing controlled analysis may be more beneficial than controlled generation.

We rank our translations and find that the 'best' translation is found within the top ten translations output by our system in the large majority of cases. We apply our *post hoc* validation method to successfully correct almost 70% of erroneous determiner-noun bigrams in our French translations.

A number of issues for future work present themselves. The test data which we use is uncontrolled. The corpus which we use as training data, although controlled, is small and as a result our system suffers from problems of data-sparseness. Although we seed our example-base with over 200,000 strings in our phrase-based EBMT system (cf. chapter 3), we would like to extend this to cover as many sententially-aligned strings.

Although our segmentation method and *post hoc* validation process reduces problems of boundary friction, the translations still suffer from errors and incorrect word order. Many of these errors can be attributed to data-sparseness. In numerous cases, no chunk matches can be found and so much of the translation is produced word-for-word. For example, the string *to modify a DNS server* is translated using the chunk and word translations in (158):

(158) <PREP> to modify ⇔ pour modifier
<DET> a ⇔ un
<LEX> DNS ⇔ DNS
<LEX> server ⇔ serveur

If the larger chunk pair <*a DNS server, un serveur DNS*> were added to the marker-lexicon, the correct translation *pour modifier un serveur DNS* could be produced. Increasing the example-base on which the system is trained should considerably lessen these problems. The weights assigned to higher quality chunk and word alignments should increase, given that they are likely to be derived more frequently from a larger training data set.

In section 4.2.2, we showed that when our revised sub-sentential alignment method was applied, 87% of sententially-aligned pairs were candidates for sub-sentential alignment, as opposed to 18% when the old algorithm was used. Moreover, we can now generate alignments of a 2:1, 3:1, 1:2, 2:2 etc. nature where previously it was only possible to produce 1:1 alignments. However, we also noted that using *Logomedia* to extract lexical correspondences has some associated drawbacks. Extracting lexical information from the corpus could be a useful development for future models. When such correspondences are extracted from a more scalable data set, they are more likely to be reliable.

Most of the problems identified occur due to the size of our corpus and/or reliance on *Logomedia*. In the following chapter, we describe how we develop a much larger scale

EBMT system. We also describe how we significantly reduce our dependency on *Logomedia*, both as a resource for producing our sub-sentential alignments and as a means of seeding our example-base.

Chapter 5

Scalability

Our novel implementation of an EBMT system based on controlled language specifications was described in chapter 4. We presented a revised algorithm which could generate subsentential alignments from example pairs when the relation between the source and target chunks was more complex than a 1:1 correspondence. In our phrase-based system (cf. chapter 3), such cases were not addressed as our algorithm was limited to producing 1:1 alignments.

We showed, using both manual and automatic evaluation metrics, how our revised algorithm contributed to an improvement in translation quality. Following a number of minor alterations to our word-level lexicon, we illustrated how we could outperform the rule-based system *Logomedia*. Consequently, we demonstrated that our results bear out the hypothesis of (Carl, 2003b; Schäler et al., 2003) that EBMT may be more suitable for controlled translation than RBMT.

At different stages, we evaluated how controlling the source and target strings affected translation performance. Although the BLEU scores obtained suggested that controlling the target language (French-English) was more rewarding, a manual evaluation and an analysis of alternative automatic metrics pointed in favour of controlling the source language (English-French). We discussed a number of reasons as to why the scores were anomalous and determined that controlling the source language has a more positive impact on translation quality than controlling the target language.

Despite these observations, our research on controlled EBMT also highlighted a number of issues. We predicted that by addressing certain key areas we could potentially improve the performance of the marker-based system.

Firstly, using an on-line MT system to construct a bitext is not an ideal approach. In chapter 4, we applied *Logomedia* for this purpose because a quality controlled bitext does not exist. The *Logomedia* system was chosen as it was deemed the 'best' on-line MT system in (Way and Gough, 2003) (cf. section 3.4.6). Secondly, despite the broad similarity between the test set and the training data, the disparity between them was extensive enough to reduce the quality of the translations produced. Moreover, any disparity was compounded by the fact that while our training data contained only just over 1,600 sentences, our test data was extracted from a TM containing over 200,000 sentences. Another factor which contributed to low quality translations was data-sparseness. While 100% coverage was observed, we noted that many of these translations were produced word-for-word, as relatively few chunk matches were located. With this in mind, we scaled up our EBMT system into a larger, more robust model (Gough and Way, 2004). In section 2.2.2 we suggested that a worthwhile experiment would involve a comparison of the performance of an SMT system and an EBMT system trained on similar data. In this chapter, we implement such a novel experiment. In the following sections we describe how we:

- scale-up our marker-based EBMT system;
- increase the similarity between our training set and our test data;
- significantly reduce our reliance on *Logomedia*;
- automatically filter incorrect alignments from our marker-lexicon;
- further improve the sub-sentential alignment algorithm described in section 4.2.1;
- provide a comparison with an SMT system using *Giza++*[1](Och and Ney, 2003).

5.1 Scalability in EBMT

(Somers, 2003) lists a number of EBMT systems in terms of the size of their example-

[1] http://www.isi.edu/~och/Giza++.html

base. The largest example-base listed is the *PanLite* system (Frederking and Brown, 1996) which contains 726,406 ⟨English, Spanish⟩ examples. The smallest is the *METLA* system (Juola, 1994, 1997) which is trained on just 7 ⟨English, Urdu⟩ example pairs.

The size of the training data in an EBMT system is dependent on the objectives of its developers. Some systems are largely experimental and may not require an extensive training set (Juola, 1994, 1997). Others, such as that of (Brown, 2003), make use of generalised templates (cf. section 2.4.4) which can help to reduce the amount of training data required by replacing similar patterns with a single general variable.

The availability of suitable corpora can also affect the size of the example-base chosen. For instance, it may be more difficult to locate bitexts for certain minority languages. One example can be drawn from Malta, where the majority of people speak both English and Maltese. However, the latter is rarely written down and therefore a parallel ⟨English, Maltese⟩ text is difficult to come by. We encountered a similar problem in our controlled EBMT system (cf. chapter 4) where a controlled bitext was unavailable to us.

In certain cases, researchers have reported an improved performance when the example-base is increased (Sumita and Iida, 1991; Sato, 1993). (Mima et al., 1998) performed an experiment where the example-base was continuously incremented. Initially, they trained the system on 100 examples and reported 30% accuracy. Gradually, they increased the example-base by incrementally adding sets of 100 examples. They report a steady increase in accuracy at each stage, with 65% accuracy reported on the final set of 774 examples. While this indicated an overall improvement of 35%, they also point out that infinitely adding more examples may not be beneficial and that there is likely to be a ceiling to this pattern.

Increasing the database can have a detrimental effect on the system if recurring examples are not dealt with appropriately. In some cases, if the same ⟨source, target⟩ example occurs more than once this can serve to reinforce the example. However, if several different target translations exist for a single source sentence then this can give rise to conflict. (Somers et al., 1994; Öz and Cicekli, 1998; Murata et al., 1999) apply a similarity metric which assigns a higher score to more frequently occurring examples. Where such a metric is not present, multiple examples can potentially result in overgeneration or ambiguity.

We calculate a weight for each translation produced by our system using the formula in (98, p.79). In our phrase-based system (cf. chapter 3), we showed how the 'best' translation was consistently ranked in the top 1% of output translations. Therefore, we do not consider recurring examples to have an adverse effect on the quality of translations produced by our system. Moreover, chunk and word alignments that occur more frequently are rewarded as they obtain a higher weight in our system.

5.2 Scalability in Marker-Based EBMT

In terms of marker-based EBMT, previous systems which apply this methodology have not been scaled up using large corpora. Prior to the implementation of our phrase-based model described in chapter 3, the largest marker-based EBMT system was the *Gaijin* system (Veale and Way, 1997) which used 1,836 ⟨German, English⟩ sentence pairs. The *METLA* system (Juola, 1994, 1997) was trained on just 29 examples for the ⟨English, French⟩ language pair and just 7 examples for ⟨English, Urdu⟩.

The TM used to seed our example-base contains 207,468 ⟨English, French⟩ sententially-aligned pairs, consisting of computer manual documentation. We used 203,529 of these sentence pairs as training data, which amounted to 4.7 million ⟨English, French⟩ words in total. The remaining 3,939 sentences were used as test data. Accordingly, we increased our example-base significantly from the previous experiment (cf. chapter 4) where only 1,691 sentence pairs were used to train the system. Our phrase-based model (cf. chapter 3), contained over 200,000 English phrases, each of which was assigned a maximum of 3 translations. However, none of these were sententially aligned pairs. This system, therefore, is larger than any other ⟨English, French⟩ system listed by (Somers, 2003) and is certainly the largest EBMT system based on the Marker Hypothesis. Table 5.1 lists marker-based EBMT systems in terms of the language pair involved and the size of the training data. The marker-based models described in this thesis are also included for comparison.

Restricting translation to a specific sublanguage domain leaves less margin for error and as EBMT systems derive translations from a set of real examples, they are perhaps more suited to sublanguage translation. We randomly extracted 3,939 sentences from the

System	*Language Pair*	*Size of Training Data*
Gaijin (Veale and Way, 1997)	English-German	1,836 sentences
METLA-1 (Juola, 1994, 1997)	English-Urdu	7 sentences
METLA-1 (Juola, 1994, 1997)	English-French	29 sentences
Phrase-based EBMT (Gough et al., 2002; Way and Gough, 2003) (cf. chapter 3)	English-French	218,697 phrases
Controlled EBMT (Gough and Way, 2003; Way and Gough, 2004) (cf. chapter 4)	English-French/French-English	1,691 sentences
Scalable EBMT (Gough and Way, 2004) (cf. chapter 5)	English-French/French-English	203,529 sentences

Table 5.1: EBMT systems which apply the Marker Hypothesis (including those presented in this thesis)

TM to use as a test set, therefore increasing the similarity between the test set and the training data and confining the system to the domain of computer manual documentation.

The remaining 203,529 sentence pairs were used as training data in our EBMT system. As in previous experiments, we applied the Marker Hypothesis to derive additional lexical resources from the sententially-aligned pairs. Given that the revised sub-sentential alignment algorithm described in section 4.2.1 contributed to an improvement in translation quality, we abandoned the naïve algorithm applied in our initial experiments (cf. section 3.2.2) and applied only the revised algorithm in our scalable system.

In chapter 4, we noted that 85% of the sententially-aligned pairs in our example-base threw up candidates for sub-sentential alignment. With our larger data set, this figure now falls to 69.7%. However, we can now deduce a set of 275,822 sub-sententially-aligned chunks. This is a considerable increase from the size of the marker-lexicon in our controlled EBMT system (cf. chapter 4), where just 6,400 sub-sentential alignments were generated when our revised algorithm was applied.

We also produced 219,406 unique generalised templates and 2,828 unique word alignments in this manner. Any words from the test set which were not present in our word-level lexicon following this alignment process were translated using *Logomedia*. This amounted to 1,993 words for English (46% of words in the test set) and 3,040 words for French (55% of words in the test set).

In chapters 3 and 4, the bitexts used to train our system were partially derived via

Logomedia. As we now have access to a large-scale sententially-aligned bitext, we restrict the use of *Logomedia* to the creation of our word-level lexicon and also use the on-line system as a baseline comparison.

In sum, we have extensively increased our training data, significantly reduced our dependence on *Logomedia* and increased the similarity between our training data and our test set. In the following section we will describe a number of experiments designed to test the impact of these steps on translation quality. We also discuss and implement some novel techniques designed to further improve the results obtained.

5.3 Experiments and Results: An Automatic Evaluation

The *Sun Microsystems* TM contains 207,468 sententially-aligned ⟨English, French⟩ pairs. 3,939 of these were randomly extracted as test data, ensuring that all words were contained within our word-level lexicon. Table 5.2 shows the minimum, maximum and average sentence lengths for the English and French test sets.

English		
Ave. Sent. Length	Min. Sent. Length	Max. Sent. Length
13.2 words	1 word	87 words
French		
Ave. Sent. Length	Min. Sent. Length	Max. Sent. Length
15.7 words	1 word	91 words

Table 5.2: Average, minimum and maximum sentence lengths for English and French test sets

We performed several experiments using our scalable model:

- translation of 3,939 test sentences;
- translation of non-exact matches;
- filtering of the marker-lexicon;
- integration of Mutual Information.

Initially we translated our 3,939 sentence test set from English-French and French-English. We performed a manual and automatic evaluation using the metrics outlined in sections 4.3 and 4.4.1. We then identified any translations in our test set which could be matched exactly with their corresponding reference translation. We eliminated these strings and evaluated the remaining translations. This enabled us to quantify the effect of exact matches on the BLEU score.

We performed a novel filtering technique, where a length-based comparison was used to identify incorrect alignments which were subsequently removed from our marker-lexicon. We assessed the overall impact of this process on translation quality.

Finally, we calculated Mutual Information (MI) scores (cf. p.152) for co-occurring source and target words and integrated these correspondences into the sub-sentential alignment algorithm. As a by-product of this process, we were able to increase the contents and improve the quality of our word-level lexicon.

In this section, we discuss the significance of our results in terms of an automatic evaluation.[2] In chapter 4 we compared our system to *Logomedia* and showed that in terms of Precision and Recall we were able to outperform the rule-based system. When the BLEU metric was used, however, *Logomedia* appeared to considerably outperform our EBMT system. Nevertheless, a manual evaluation and a closer inspection of the BLEU score indicated that there was perhaps less disparity between the two systems and that the BLEU metric is indeed quite harsh in its measurement of translation quality.

In this section, we assess the effects of increasing our training data on translation quality using automatic evaluation metrics. As in chapter 4, we use the NIST MT Evaluation Toolkit to calculate BLEU scores (Papineni et al., 2002) and we derive figures for Precision and Recall using the tools outlined in (Turian et al., 2003). We also calculate the word and sentence error rates (WER and SER) as previously.

5.3.1 Evaluation of the Complete Test Set

Table 5.3 shows the BLEU scores obtained when our 3,939 sentence test set is translated from English-French and French-English. The BLEU scores obtained by *Logomedia*[3] on

[2] The results of our evaluation differ from those in (Gough and Way, 2003) due to bug-fixes within our system.

[3] When *Logomedia* was used to translate the test set in our controlled system (cf. chapter 4), we noted that in terms of an automatic evaluation, the on-line system performed better from English-French. When our scalable test set is translated using *Logomedia*, we find that the French-English translations produced are better. Further tests would be necessary to determine the reason for this trend. However, as *Logomedia*

the same test set are included for comparison. We also show the highest BLEU score obtained by our controlled EBMT system.

System (En-Fr)	*BLEU*		*System (Fr-En)*	BLEU
Scalable EBMT	0.3040		Scalable EBMT	0.3314
Logomedia	0.1229		Logomedia	0.1313
Controlled EBMT	0.1352		Controlled EBMT	0.1703

Table 5.3: Comparing our scalable EBMT system with *Logomedia* and our controlled EBMT system using the IBM BLEU automatic evaluation metric on a 3,939 sentence test set

In our English-French system, we outperform *Logomedia* by over 147% and our controlled EBMT system by over 124%. When translating from French-English, we outperform *Logomedia* by over 152% and our controlled system by over 94%.

System	*Precision*	*Recall*	*WER*	*SER*
Scalable EBMT *(En-Fr)*	0.4772	0.6029	83.0	89.2
Logomedia	0.4190	0.4321	89.7	97.8
Controlled EBMT	0.3891	0.5293	64.8	84.0
Scalable EBMT *(Fr-En)*	0.5421	0.6709	76.8	71.3
Logomedia	0.4591	0.5534	89.7	96.1
Controlled EBMT	0.3005	0.3646	80.1	88.0

Table 5.4: Comparing our scalable EBMT system with *Logomedia* and our controlled EBMT system using automatic evaluation metrics on a 3,939 sentence test set

Table 5.4 shows the figures for Precision and Recall and WER/SER for our scalable system. Again, we include the figures for our controlled EBMT system and *Logomedia* for comparison.

We outperform *Logomedia* for both language directions in terms of Precision and Recall. For French-English translations we obtain a score 18% higher for Precision and 21% higher for Recall. For English-French translations Precision and Recall are approximately 14% and 39% higher for our system. We also obtain better scores with regard to WER and SER than the on-line system. For French-English our WER is 76.8% compared to 80.1% for *Logomedia*. For English-French the WER for our system is 83% compared to 89.7% for *Logomedia*. As regards SER, we score 71.3% for French-English (*Logomedia* obtains 96.1%). When translation from English-French our SER is higher at 89.2% but remains

is used only as a baseline comparison against our system, we deem such an experiment unnecessary for the purpose of this thesis.

better than the score for *Logomedia* (97.8%).

Therefore, taking all automatic evaluation metrics into consideration, we significantly outperform *Logomedia*. This is most likely due to the fact that *Logomedia* is a general-purpose system and as a result it does not have recourse to all the domain-specific vocabulary present in our training data. The increased similarity between our training data and our test set is also responsible for the improvement in these figures.

We outperform our controlled EBMT system in terms of Precision and Recall for both languages. For French-English, our figure for Precision is over 80% higher in our scalable system and Recall is over 84% higher. For English-French translations, Precision improves by over 22% from our controlled EBMT system and Recall increases by almost 14%. In our French-English system we obtain better figures for WER (76.8% compared to 80.1%) and SER (71.3% compared to 88%). However, where control was exerted over the source language (English-French) in our controlled EBMT system, the WER and SER for our scalable system deteriorates. Our controlled system obtained a WER of 64.8% and a SER of 84% when translation was performed from English-French. In our scalable system, the figures for WER and SER for English-French are 83% and 89.2% respectively.

These figures for our controlled system were obtained by improving our word-level lexicon and making some minimal adjustments to the derivation of our translations (p. 123). Currently, we seed our word-level lexicon by aligning the content and marker words from bigrams produced via the Marker Hypothesis (cf. section 3.2.2). We also rely on *Logomedia* to generate a translation for any words in our test set which have not been included in our word-level lexicon following this procedure. In section 5.3.4, we will show how integrating MI to derive additional word alignments improves these figures for our scalable system.

In section 4.4 the BLEU scores indicated that performing translation for French-English yielded better results. Other automatic evaluation metrics conflicted with this pattern and suggested that performing translation from English-French produced better results. In our scalable system, all the automatic evaluation metrics corroborate one another and all suggest that our system is subject to a better performance when translating from French-English. As we pointed out in section 4.4.4, this is to be expected given that the

French language is more morphologically rich than English and translating from English to French gives rise to more problems of boundary friction than translating into English. In our controlled EBMT system, we attributed the superior quality French translations to the beneficial effects of controlling the source strings. Given that our automatic evaluation metrics now suggest a better performance from French-English, this may be further evidence in favour of controlling the source language.

5.3.2 Evaluation of Non-Exact Matches

We predicted that training our system on a specific domain and restricting our test set to a similar area would improve the results obtained in an automatic evaluation. The figures presented in Tables 5.3 and 5.4 confirm this to be the case.

Each target language translation is submitted for evaluation along with its reference translation i.e the oracle or gold standard translation for the original source language string. If the reference translation and the target language translation produced via EBMT are identical then a BLEU score of 1 is returned for that translation. In order to assess the effect of exact matches on the BLEU score, we identified those translations for which an exact reference match could be located and eliminated the associated source sentences from our test set.

For our French translations, this amounted to 426 strings, while 1,130 of our English translations obtained an exact match with a reference translation. As a baseline comparison, we also calculated the BLEU score obtained when these sentences were translated via *Logomedia*. In addition, we calculated figures for Precision, Recall and WER/SER for the non-exact test set. The results are shown in Table 5.5.

System	BLEU	Precision	Recall	WER	SER
Scalable EBMT *(En-Fr)*	0.2588	0.4499	0.5722	93.0	100
Logomedia	0.1163	0.3962	0.4029	93.9	99.3
Scalable EBMT *(Fr-En)*	0.2717	0.5193	0.6785	84.4	100
Logomedia	0.1639	0.4779	0.5722	91.8	98.6

Table 5.5: Comparing our scalable EBMT system (exact matches eliminated) with *Logomedia* using Automatic Evaluation metrics

When our entire test set was evaluated for English-French, we outperformed *Logomedia* by over 147%. When the exact reference matches are eliminated, we would expect this figure to fall. However, we continue to outperform *Logomedia* by over 122% in terms of BLEU when the test set is confined to non-exact matches. In terms of Precision and Recall, we now outperform *Logomedia* by over 13% and 42% respectively. Not surprisingly, the automatic metrics suggest that the non-exact matches are lower quality. However, while the BLEU score decreases by 17% for the English-French translations in our complete test set, Precision and Recall fall by only 6% and 5% respectively. The 100% sentence error rate obtained by our system is to be expected, given that none of these sentences can be matched exactly with a reference translation. WER rises from 83% to 93%.

For our French-English system, we outperform *Logomedia* in terms of BLEU score by over 65%. With regard to Precision, our system obtains a score 8% higher and for Recall a score 18% higher. We obtain a WER of 84.4%, compared to 91.8% for *Logomedia*. When the entire test set was evaluated we outperformed the online system by 152% with regard to BLEU score, 18% for Precision and 21% for Recall.

Unsurprisingly, eliminating the exact-matches from our test set adversely affects the automatic evaluation figures. In section 4.4.4, we observed that the BLEU score can sometimes be unduly harsh when evaluating translations which are not very similar to the reference sentence. When our non-exact test set is evaluated, we find that BLEU penalises the translations more extensively than the alternative metrics. However, we find that we continue to significantly outperform *Logomedia* and produce high quality translations.

In the following sections, we describe some additional experiments. We filter the data in our marker-lexicon and examine what repercussions this has on the quality of our French translations. We also integrate MI and measure its contribution to translation performance.

5.3.3 Filtering the Data

In section 4.2.2 we mentioned that a possible avenue for future work might be to filter some of our incorrect alignments in a pre-processing stage. In this section, we describe how we implement such a technique and assess the overall effects of the filtering process on translation quality.

For each chunk in our marker-lexicon derived via the Marker Hypothesis, we produce an alternative translation using *Logomedia.* We then perform a length-based comparison between the target chunk in our lexicon and the translation produced by *Logomedia.* For example, the incorrect alignment in (159) is present in our marker-lexicon.

(159) your password : votre mot

When the source string *your password* is translated via *Logomedia*, the translation produced is *votre mot de passe.* As this translation differs by more than 1 word from the target string *votre mot* in our alignment, the incorrect alignment in (159) derived via the Marker Hypothesis is effectively 'filtered out' of our marker-lexicon.

Integrating this novel filtering technique causes a reduction in the size of our lexical resources. We note that of the 275,822 sub-sententially-aligned chunks derived via the Marker Hypothesis, 141,070 are eliminated. This is a loss of almost 51%. The number of unique word alignments falls by 10.5% to 2,531. Likewise, the number of generalised templates is reduced to 110,581 from 219,406 — a loss of almost 50%. The filtered resources were used to seed the memories of our system and the 3,939 sentence test set was re-submitted for translation from English to French. The results obtained are in Table 5.6.

System	*BLEU*	*Precision*	*Recall*	*WER*	*SER*
Scalable EBMT (original)	0.3040	0.4772	0.6029	83.0	89.2
Scalable EBMT (filtered)	0.4040	0.5953	0.6999	59.8	86.5
Logomedia	0.1229	0.4190	0.4321	89.7	97.8

Table 5.6: Comparing our scalable EBMT system (English-French: filtered data) with *Logomedia* using Automatic Evaluation metrics on a 3,939 sentence test set

All of the automatic metrics reflect an improvement in translation quality when the training data is filtered. The BLEU score for our test set improves by 33% from the original figure. Precision and Recall also improve by 25% and 16% respectively. The WER improves by 39% and the SER by 3%.

We now outperform *Logomedia* by 229% in terms of BLEU score. The percentage improvement over *Logomedia* for Precision and Recall also improves when the data is filtered (42% and 62%). Therefore, despite the reduction in our lexical resources, we can conclude that the filtering process has a positive effect on translation quality.

5.3.4 Integrating Mutual Information

In section 4.2.1, we noted that including a baseline dictionary to derive lexical correspondences between source and target chunks increased the number of sub-sententially-aligned fragments derived from the corpus. The BLEU score for our translations also increased by 3% for our English translations and more significantly by 44% for our French translations. A dictionary with 100% coverage was produced very efficiently using *Logomedia* to translate all words in our corpus.

However, given the general-purpose nature of the on-line rule-based system, we encountered several instances where the dictionary failed to be a useful resource for deriving lexical equivalences. In these cases, the word translation produced via *Logomedia* differed from the actual word translation in our corpus. For example, in section 4.2.2 we noted that the translation of *hide* by *Logomedia* is the noun *peau.* The translation of the word *hide* in our corpus, however, is mainly one of two verb forms, *masquer* or *cacher.*

By deriving our correspondences using MI scores, we can add word correspondences such as ⟨hide,cacher⟩, ⟨hide,masquer⟩ to our word-level lexicon and align the chunks containing these word correspondences. Consider the example in (160):

(160) <NULL> use <DET> the permissions folders <PREP> to restrict access ⇔ <NULL> utilisation <PREP> des dossiers <PREP> d'autorisations <PREP> pour restreindre <DET> les accès

When we retrieve the translations for the content source words from our baseline dictionary, we obtain the correspondences in (161):

(161) a. use ⇔ utilisation

b. permissions ⇔ permissions

c. folders ⇔ classeurs

d. restrict ⇔ restreignez

e. access ⇔ accès

We can use the correspondences in (161b) and (161e) to create lexical equivalences between the source and target chunks in (160). However, the source words *use*, *folders* and *restrict* in (161a), (161c) and (161d) cannot be linked to any target words in the associated translation when we rely solely on *Logomedia* for this purpose.

We overcome this problem by extracting additional correspondences directly from the corpus. In a pre-processing stage, we calculate MI scores for co-occurring ⟨source, target⟩ words in our corpus using the formula in (162) (Church and Hanks, 1990).

(162) $$MI(x, y) = log_2 \frac{P(x,y)}{P(x)P(y)} = log_2 \frac{Nf(x,y)}{f(x)f(y)}$$

The probabilities *P(x)* and *P(y)* are calculated by counting the total number of occurrences of a source word *f(x)* and a target word *f(y)* in a corpus. These figures are normalised by the size of the corpus (N). The joint probability *P(x,y)* is estimated by counting the number of instances where a source and target word occur within the same sententially-aligned pair *f(x,y)* and normalising this figure by the size of the corpus.

We assume that ⟨source, target⟩ words that frequently occur in the same example pair are more likely to be translations of one another than those ⟨source, target⟩ words that occur together less frequently or not at all. A high MI score suggests that a ⟨source, target⟩ pair co-occur frequently. A score < 0 indicates that two words do not co-occur within the same ⟨source, target⟩ pair. A low score suggests that a ⟨source, target⟩ pair occur together infrequently.

At the sub-sentential alignment stage, MI figures are extracted for each source word that has not been linked to a word in the target string using our baseline dictionary. Table 5.7 shows the top-three MI scores calculated for the relevant source words in (160). Note that marker words such as *the* and *to* are considered stop words and therefore MI scores are not calculated for these lexical items. In any case, we consider that these words are already factored into the alignment process via the marker tags.

We note that *use* co-occurs most frequently with *utiliser.* However, given that the word *utiliser* is not present in the current target string, this equivalence is not considered. The same applies to its second most frequently co-occurring target word *utilisez.* Following the exclusion of these two potential correspondences, a link can be drawn between *use* and its third most frequently co-occurring word *utilisation*, which does appear in the target string.

Source word	Target word	MI score
folders	dossiers	13.558771583624
folders	restreindre	10.7853817469554
folders	sous-dossiers	10.286978643188
use	utiliser	12.2831687936086
use	utilisez	12.2052631365424
use	utilisation	9.61244467887597
restrict	limiter	12.222319840894
restrict	restreindre	11.6425013456411
restrict	id-connexion	10.4568776799834

Table 5.7: Top-three MI scores for some source words which have not been linked using the baseline dictionary derived via *Logomedia*

Similarly, the source word *restrict* co-occurs most frequently with a target word, *limiter*, which is not present in the current target string. In a similar procedure to that used to derive the correspondence between *use* and *utilisation*, *restrict* is linked to the target word *restreindre*.

A constraint is also imposed which determines that for a correspondence to be formed between a source and target word, the MI score for that word pair must be higher than the MI score for any alternative word pair which includes that target word. This is best explained by referring to the example in (160). When *limiter*, the primary candidate for forming a correspondence with the source word *restrict* is eliminated, the target word *restreindre* occurs most frequently with *restrict*. Although *restreindre* also co-occurs with the source word *folders*, the MI score for its co-occurrence with *restrict* is higher (11.6 compared to 10.7) and therefore *restrict* and *restreindre* are linked. The source word *folders* co-occurs most highly with *dossiers* which is also present in the target string and therefore these ⟨source, target⟩ words can be linked.

Following the integration of the MI scores, the lexical equivalences in (163) are generated:

(163) use ⇔ utilisation
folders ⇔ dossiers
restrict ⇔ restreindre

Along with the equivalences produced using the baseline dictionary, these are applied to derive the sub-sententially-aligned chunks in (164) and the generalised templates in

(165):

(164) <NULL> use ⇔ utilisation
<DET> the permissions folders ⇔ des dossiers d'autorisations
<PREP> to restrict access ⇔ pour restreindre les accès

(165) <DET> permissions folders ⇔ <PREP> dossiers d'autorisations
<PREP> restrict access ⇔ <PREP> restreindre les accès

When MI is integrated into the sub-sentential alignment process, the number of unique aligned chunks increases by just over 6%. However, the quality of our alignments also improves and alignments such as those of a 2:3 nature can be correctly derived where previously they contained errors. For instance, in the segmented ⟨source, target⟩ pair in (166) the first two chunks in the source sentence are aligned with the first two chunks in the target sentence.

(166) <PPRON>$_{s1}$ you use <DET>$_{s2}$ the left page layout option
<PREP>$_{s3}$ to apply <DET>$_{s4}$ the page layout settings ⇔
<DET>$_{t1}$ l'option <PREP>$_{t2}$ de mise en page <PREP>$_{t3}$ à gauche sert
<PREP>$_{t4}$ à définir <PREP>$_{t5}$ des paramèteres <PREP>$_{t6}$ de mise en page

Prior to the integration of MI this would not have been possible as the lexical equivalence between *use* and *sert* would not have been established. In addition, the 1:2 alignment between chunk *s4* and chunks *t5* and *t6* would not have been derived because *settings* could not be linked to *paramèteres* using the base dictionary. When MI is integrated however, the alignments in (167) are produced:

(167) <PPRON> you use the left page layout option ⇔ l'option de mise en page à gauche sert
<PREP> to apply ⇔ à définir
<DET> the page layout settings ⇔ des paramèteres de mise en page

The lexical equivalences which are derived during this process are also added to the word-level lexicon. For example, from the sentence pair in (160), the word-level lexicon is assigned the equivalences in (163). Applying this process means that our word-level lexicon increases by over 423%. This ultimately reduces the number of words which are translated using *Logomedia*. Previously, we relied on *Logomedia* to translate 46% of the words in our English test set (1,993 words in total). Following the integration of MI, we now use the on-line system to translate only 19% or 814 of the words in our English test set. Similarly, the number of words in our French test set which need to be translated via *Logomedia* falls from 55% (3,040 words) to 20% (1,119 words).

A complete summary of the results obtained when MI scores are integrated is given in Table 5.8.

System (En-Fr)	**BLEU**	**Precision**	**Recall**	**WER**	**SER**
Our System (original)	0.3040	0.4772	0.6029	83	89.2
Our System (filtered)	0.4040	0.5953	0.6999	59.8	86.5
Our System (MI)	0.4409	0.6727	0.6877	52.4	65.6
Logomedia	0.1229	0.4190	0.4321	89.7	97.8

System (Fr-En)	**BLEU**	**Precision**	**Recall**	**WER**	**SER**
Our System (original)	0.3314	0.5421	0.6709	76.8	71.3
Our System (MI)	0.4611	0.6782	0.7441	50.8	51.2
Logomedia	0.1313	0.4591	0.5534	89.7	96.1

Table 5.8: Comparing our EBMT system to *Logomedia* using Automatic Evaluation Metrics: (integrating MI)

We can see from the figures in Table 5.8 that integrating MI leads to an improvement in translation quality for both language directions. For English-French, the BLEU score improves by 45% from the baseline figure in our original scalable model. Precision and Recall improve by 41% and 14% respectively. WER improves by 58% and SER by 36%.

For French-English, the BLEU score improves by 39%, Precision by 25% and Recall by almost 11%. WER and SER improve by 51% and 39%. We observe therefore, that including MI results in an improvement in translation quality for both English-French and French-English translations. Not surprisingly, the latter continues to obtain better results. However, the disparity between the scores for the English and French translations is also reduced. In our baseline model, French-English translations obtained an average BLEU score 9% higher than English-French translations. When we integrate MI, this figure falls to 4%. Similarly, the percentage difference for Precision falls significantly from 12% to just 0.9%. The disparity between English and French translations with regard to Recall and WER is also reduced.

It would appear, therefore, that the integration of MI is beneficial when translating from both French-English and English-French. These benefits seem to be heightened when translating from English-French. This may be due to the ability of MI to capture some equivalences for different morphological variants in French.

5.4 Experiments and Results: A Manual Evaluation

We also carried out a manual evaluation using the notions of intelligibility and accuracy (cf. p.123). Our objective was to confirm the findings of the automatic evaluation and to provide a closer inspection of the effect of integrating our filtering technique and including MI. For our English and French translations, we extracted 100 sentences at random from the portion of our test set for which an exact match could not be found (cf. section 5.3.2). The translations were evaluated by a native English speaker with good French competence. The translations produced following the filtering of the marker-lexicon and the integration of MI were also examined for comparison.

5.4.1 Evaluating English-French Translations

Tables 5.9 and 5.10 present the results of carrying out a manual evaluation on our English-French translations. The results for intelligibility are given in (5.9).

System	Score 0	1	2	3
Our System (original)	16	36	33	15
Our System (filtered)	15	36	32	17
Our System (MI)	5	10	20	65

Table 5.9: A Manual Evaluation of our EBMT system (English-French): Intelligibility

System	Score 0	1	2	3	4
Our System (original)	12	14	23	32	19
Our System (filtered)	12	14	21	29	24
Our System (MI)	4	6	20	19	51

Table 5.10: A Manual Evaluation of our EBMT system (English-French): Accuracy

The manual evaluation suggests that the filtering technique only marginally improves the accuracy and intelligibility of our translations. The number of translations obtaining a score of 3 or 4 for accuracy increases by 2% from our original model, while the number of translations obtaining a score of 2 or 3 for intelligibility increases by only 1%. When we compare this to our automatic evaluation, we observe that although the overall SER only improved by 3%, WER improved by 39%. Furthermore, the BLEU metric indicated a 33% increase over the original model, while Precision improved by 25%.

Given that our automatic evaluation was performed on the entire test set (3,939 sentences) and our manual evaluation on a set of 100 randomly extracted sentences, some level of disparity can be expected. However, while our automatic evaluation strongly implies that filtering the data improves translation quality, our manual evaluation suggest that it is far less influential. In section 5.5 we provide some further insights into this disparity.

When the system is seeded with the resources generated using MI, the percentage of translations that obtain a score of 2 or 3 for intelligibility rises from 48% to 85%. This 37% improvement is more comparable with our automatic evaluation, where following the integration of MI, the BLEU score increased by 45% for English-French translations. Precision and Recall increased by 41% and 14% respectively. The figures for accuracy in our manual evaluation suggest an improvement of 19% when MI is included. Therefore, according to an automatic and manual evaluation, the inclusion of MI considerably improves the quality of our translations for English-French.

5.4.2 Evaluating French-English Translations

We carried out a similar manual evaluation on 100 of the English translations obtained. We did not perform French-English translation using the filtered data and therefore the improvement from the original model to the integration of MI in our system was measured.

System	Score 0	1	2	3
Our System (original)	15	33	31	21
Our System (MI)	4	11	22	63

Table 5.11: A Manual Evaluation of our EBMT system (French-English): Intelligibility

System	Score 0	1	2	3	4
Our System (original)	7	10	20	38	25
Our System (MI)	2	6	11	45	36

Table 5.12: A Manual Evaluation of our EBMT system (French-English): Accuracy

When MI is applied, we observe that intelligibility (score 2 and 3) improves for French-English translations by 33% and accuracy improves by 18%.

In an automatic evaluation, translating in the direction French-English produced consistently better results than translating from English-French, even when MI was integrated. The results of the manual evaluation indicate that French-English translations are more intelligible than English-French translations when the original model is used. A higher accuracy for English-French translations is also reported following the integration of MI. 81% of translations obtain a score of 3 or 4 compared to 70% of French translations. However, when the same translations are evaluated in terms of intelligibility, the English and French translations examined are deemed equally intelligible and the number of translations obtaining a score of 2 or 3 is 65% for both English and French.

5.5 Discussion of Results

The results of the manual evaluation support those provided by the automatic evaluation and suggest that our results improve when the filtering technique is applied to our marker-lexicon. They also provide strong evidence in favour of using MI to upgrade the quality of our translations. To illustrate how the filtering mechanism and the integration of MI improve our results, consider the English input string in (168):

(168) To obtain some additional information about sun cluster see the web site http://www.sun.com/clusters

The reference translation for the input in (168) is shown in (169):

(169) Pour de plus amples renseignments concernant sun cluster consultez le site web http://www.sun.com/clusters

When the original scalable model is used, the highest scoring translation is that shown in (170):

(170) Pour de plus amples renseignments cliquez propos de sun grappe voir le site web http://www.sun.port/grappe

The translation in (170) shares the n-gram *pour de plus amples renseignments* with the reference translation in (169). When evaluated using automatic metrics, the 4-grams, *pour de plus amples* and *de plus amples renseignments* are rewarded and the translation receives a BLEU score of 0.5868. However, the translation of the source n-gram *to obtain some additional information* has been retrieved from the marker-lexicon as *pour de plus amples renseignments cliquez.* Furthermore, the translations of the words *com* and *clusters* are incorrect in this instance and do not match with the reference translation.

When the translation of the same sentences is produced using the filtered data, the highest weighted output string is that in (171):

(171) Pour de plus amples renseignments de sun cluster voir le site web http://www.sun.port/les grappe

The chunk *pour de plus amples renseignments cliquez* has been filtered from the marker-lexicon as a translation of *to obtain additional information.* The BLEU score returned for the translation in (171) is 0.6334, an improvement of almost 8% from the translation produced when the unfiltered data is used.

When MI is integrated, the translation in (172) is derived. The translation of *cluster* is correct in this instance as *cluster* and *clusters* translates as *clusters.* It also translates *see* as *consultez*, and *com* as *com*, both of which are contained in the reference translation. The BLEU score for this translation is 0.7674, an improvement of over 21% from when the filtered data is used and an improvement of almost 31% from when the the baseline model is used.

(172) Pour de plus amples renseigments sur sun cluster consultez le site web http://www.sun.com/clusters

In an automatic evaluation, we noted that filtering the marker-lexicon improved the quality of the translations produced. A manual evaluation agrees with this trend. However, the results are far less significant than in our automatic evaluation.

Our EBMT system has the ability to output numerous translations. As in our controlled EBMT system (cf. chapter 4), we performed an on-the-fly pruning of our translations. This process eliminates lower weighted chunks dynamically so that only the top 200 translations are retained at any one time. In any case, only the highest weighted translation is submitted for automatic evaluation. However, in a manual evaluation, we noted that when using the original model, the 'best' translation was ranked in first position for only 31% of cases. When the filtered data is used, however, the 'best' translation can be located first in 88% of cases. As the top-ranked translation is submitted for automatic evaluation, it is likely that this had an impact on the results of the automatic evaluation.

In some cases, filtering certainly improves the quality of alignments in our marker-lexicon and improves the quality of our translations (cf. (171)). In some instances, by filtering target chunks, better alignments are more likely to be ranked higher by our system. For example, the source chunk *the benefits* has ten possible translations in our marker-lexicon. These all occur once and thus are are each assigned a weight of $\frac{1}{10}$. Therefore, the correct translations *les avantages* and *les bénéfices* are not rewarded by our system and there is less likelihood that they will be contained in the highest ranking translation output by the system. When the marker-lexicon is filtered, however, the incorrect translations of *the benefits* are eliminated and the possibility of either *les avantages* or *les bénéfices* occurring as the top-ranked translation increases significantly.

However, filtering the data does have side-effects. We observe that some correct alignments are removed from the marker-lexicon using this process. For example, the well-formed chunk in (173) is removed according to the length-based criterion:

(173) the log files are stored ⇔ les fichiers journaux sont stockés

When MI is included in our system, the chunk in (173) is retained. We also observe that the overall quality of our translations improves according to both an automatic and manual evaluation. The string *les avantages* now occurs ten times in our marker-lexicon and is the highest weighted translation of *the benefits.* Without filtering the data in this instance, *les avantages* will more than likely appear in a top-ranked translation of a sentence containing *the benefits.* Furthermore, we note that in 90% of cases, the best translation is ranked first when MI is used.

Given that integrating MI improves the quality of translations further and significantly increases and improves the contents of our word-level lexicon, it may not be that filtering the data is the most effective way of improving translation quality. Perhaps scaling up the data and improving our resources in this way is more beneficial and increases the likelihood that the higher quality data in our lexicons will outweigh the poorer quality data.

5.6 Comparing our marker-based EBMT system to an SMT system

SMT systems (p. 20) derive a language model and a translation model from a (usually very large-scale) bilingual corpus. Candidate translations are produced by maximising the probabilities in the language and translation models. Currently SMT is the more popular research paradigm and EBMT, despite its advantages over SMT (p. 20), appears in some respects to be less prominent in novel MT experiments.

Although some recent research in SMT has recognised the benefits of including syntactic information and phrasal correspondences, these features had been present in the earliest EBMT systems. In section 2.2.2 we hypothesised that EBMT could outperform SMT when the same training data is used and the size of the training data is reasonably scalable. We considered that a useful and novel experiment would involve comparing an EBMT system based on the *linguistics-lite* approach described in this thesis to an SMT system which integrates word correspondences (Way and Gough, forthcoming).

To this end, we used the following tools to develop a language and translation model for an SMT system:

- *Giza++*(Och and Ney, 2003)[4];
- the CMU-Cambridge statistical toolkit[5];
- the ISI ReWrite Decoder[6].

We randomly extracted a 3,939-sentence testset from the original 207,468-sentence *Sun Microsystems* TM. The remaining 203,529 sentences were used as training data, split three ways:

- Training Set 1 (TS1): 50,882 English-French sentence pairs;
- Training Set 2 (TS2): 101,765 English-French sentence pairs (inc. TS1);
- Training Set 3 (TS3): 203,529 English-French sentence pairs (inc. TS1 and TS2).

Table 5.13 shows the results obtained for English-French for both our EBMT system and the SMT system.

		Bleu	**Precision**	**Recall**	**WER**	**SER**
TS1	SMT	0.2971	0.6739	0.5912	54.9	90.8
	EBMT	0.3318	0.6525	0.6183	54.3	89.2
TS2	SMT	0.3375	0.6824	0.5962	51.1	89.9
	EBMT	0.4534	0.7355	0.6983	44.8	77.5
TS3	SMT	0.3223	0.6513	0.5704	53.5	89.1
	EBMT	0.4409	0.6727	0.6877	52.4	65.6

Table 5.13: Comparing our EBMT system with an SMT system trained on the same data using Automatic Evaluation Metrics: English-French

In the first training set, SMT outperforms EBMT in terms of Precision (0.6739 compared to 0.6525). However, for the most part, the automatic evaluation metrics suggest that EBMT can outperform SMT from English-French. When the system is augmented with additional training data, the Bleu score suggests that the EBMT system incrementally improves. With the exception of SER however, the remaining metrics suggest that

[4] http://www.isi.edu/~och/Giza++.html

[5] http://mi.eng.cam.ac.uk/ prcl4/toolkit.html

[6] http://www.isi.edu/licensed-sw/rewrite-decoder/

when the system is trained on just over 100,000 sentence pairs it yields better results than when it is trained on just over 200,000 sentences. It is possible that a degree of overfitting may affect these results. However, it is also possible that the weights assigned to our translations offer a reason for this trend. While our EBMT system can output numerous translations for a given sentence, only the top-ranked translation is then submitted for automatic evaluation. However, the top-ranked translation is not necessarily the 'best' translation. It may be the case, therefore, that increasing the training data does not result in a deterioration in translation quality but does adversely affect the weighting of translations.

With regard to SMT, the automatic evaluation metrics suggest an improvement from training data 1 to training data 2. The figure for SER also improves from training data 2 to training data 3. However, given that Precision, Recall and WER suggest a drop in translation quality when the system is trained on 203,529 sentences, overfitting may also be an issue here. It seems that when the system is trained on just over 100,000 sentences, optimal results are obtained for both SMT and EBMT for this particular test set, in terms of Bleu score, Precision, Recall and WER. As it is generally assumed that increasing the training data in an SMT system will improve the quality of the output translations, these results are particularly surprising.

Although, we observe that the figures for SER improve for both SMT and EBMT from T1 to T3, we note that the improvement for EBMT is more significant (26% compared to 0.1%). This is likely to be the result of an increase in the number of exact matches located when the training data is increased. When the system is trained on 203,529 sentences (TS3), the number of test sentences which can be retrieved directly from our example-base is approximately 10%. This means that for the 34.4% of translations which obtain a 0% SER score, 24.4% of these are produced using chunks derived from the Marker Hypothesis.

The results for French-English translations are presented in Table 5.14.

All the automatic evaluation metrics show that the SMT model obtains better results for French-English. The EBMT system also produces better translations from French-English in terms of Bleu, Recall and SER. However, with regard to WER, the

		Bleu	**Precision**	**Recall**	**WER**	**SER**
TS1	SMT	0.3794	0.7096	0.7355	52.5	86.5
	EBMT	0.2571	0.5419	0.6314	69.7	89.2
TS2	SMT	0.3924	0.7206	0.7433	46.2	81.3
	EBMT	0.4262	0.6731	0.7962	55.2	66.2
TS3	SMT	0.4462	0.7035	0.7240	46.8	80.8
	EBMT	0.4611	0.6782	0.7441	50.8	51.2

Table 5.14: Comparing our EBMT system with an SMT system trained on the same data using Automatic Evaluation Metrics: French-English

French-English translations obtain an optimal score of 50.8% compared to a better score of 44.8% for English-French. For Precision, the EBMT system obtains a maximum score of 0.7355 for English-French but only 0.6782 for French-English. Intuitively, translating from French-English should yield better results as there are less problems concerning agreement errors and boundary friction. However, as there are potentially several translations for a single word, it is possible that the top-ranked translation submitted for evaluation may contain words which, while accurate and intelligible, do not correspond to those in the reference translation.

Although the metrics obtained for English-French strongly suggested that EBMT outperforms SMT, the results for French-English are not as conclusive. On the initial TS1 set from French-English, EBMT does not outperform SMT for *any* of the five metrics. An improvement is noted when the system is trained on just over 100,000 sentences (TS2). EBMT now outperforms SMT in terms of Bleu score, Recall and SER (66.5% compared to 81.3% for SMT). However, SMT still produces better translations according to Precision and WER (46.2% compared to 55.2%). This trend continues on the final training set (TS3). SMT continues to outperform EBMT in terms of Precision and WER, albeit less significantly (3.7%), but EBMT wins out according to the remaining metrics.

Our results show, that both EBMT and SMT can produce better results for French-English translation compared to English-French. Of the five automatic evaluation metrics for each of the three training sets, in nine of the fifteen cases SMT wins out over our EBMT system. However, when compared against SMT, the results for our English-French system are much more significant as fourteen of the fifteen scores indicate that EBMT can outperform SMT. Therefore, in summary, EBMT outperforms SMT in 20 tests, while

SMT does better in 10 experiments. Ultimately, EBMT can be seen to outperform SMT by a factor of two to one.

5.7 Discussion

In this chapter we have presented a scalable EBMT system based on the Marker Hypothesis. As far as we are aware, this is the largest English-French system in existence and certainly the largest marker-based EBMT system which is trained on sententially-aligned pairs. We have significantly reduced our dependence on the on-line rule-based system *Logomedia* from previous experiments (cf. chapters 3 and 4). We have shown, using manual and automatic evaluation metrics, that when the revised sub-sentential alignment algorithm (cf. section 4.2.1) is applied to a set of 203,529 aligned sentences, the quality of our translations improves. We demonstrated that when similar test data is used, we can significantly outperform *Logomedia* and illustrated that for the majority of automatic evaluation metrics our results improve from our controlled EBMT system (cf. chapter 4). In addition, we have provided a novel comparison between an SMT system and an EBMT system using the *Giza++* tool and have shown that EBMT can outperform SMT by a factor of two to one.

We have presented various experiments to test our EBMT system and have investigated different ways to improve the results obtained. When the data in the marker-lexicon was automatically filtered, our automatic evaluation metrics suggested that the quality of our translations improved. A manual evaluation also indicated that filtering the data in our marker-lexicon improved translation performance but much less significantly. We observed that filtering the data certainly improved our resources and raised the 'best' translation to a higher position in our set of ranked candidate translations. However, we also noted that filtering the data can adversely affect the system.

We used MI to further improve our sub-sentential alignment algorithm and this increased the size of our word-level lexicon significantly. We observed that the quality of our alignments improved. When our system was seeded with the resources derived using MI, the quality of our translations also improved and the number of 'best' translations ranked first increased.

We concluded that scaling up good quality data and extracting information from the corpus may be more beneficial than eliminating data by a filtering process. Future work could involve further analysis on the positive and adverse affects of the filtering process. One issue to address might be the weights assigned to translations derived from our lexical resources. Currently, translations derived from our word-level lexicon, generalised-lexicon and marker-lexicon are weighted equally. We could prioritise the lexicons so that more weighting is assigned to chunks derived from the marker-lexicon than to those derived via word insertion using chunks from the generalised-lexicon and word-level lexicon. One possible means of initiating this process might be via the Weighted Majority Algorithm (Littlestone and Warmuth, 1992).

Performing translation from French-English gives better results than from English-French. However, when MI is integrated the results are more comparable in both an automatic and a manual evaluation. We note that the English translations obtain a higher accuracy. This is mainly due to the heightened problem of boundary friction when translating from English-French. However, when intelligibility is measured, the French and English translations are more comparable. MI benefits both English and French translations but has a more positive impact on the latter. As was noted in section 4.2.2, allowing MI to match source words with more than one target word could further improve the performance of the system.

Despite the fact that SMT is arguably the more popular research paradigm, we have shown that even a *linguistics-lite* EBMT system can outperform an SMT system which uses just word correspondences. This provides further evidence of the usefulness of syntax and phrase correspondences in a corpus-based system and shows that SMT will not inevitably outperform EBMT.

Chapter 6

Conclusions

The 'knowledge acquisition bottleneck' has proved to be the major stumbling block for rule-based approaches to MT. The manual development of large-scale grammars and rules is time-consuming, expensive and prone to error. Corpus-based approaches do not generally require hand-crafted rules and extensive grammars and, in this respect, can provide a solution to the problems of knowledge acquisition.

Corpus-based approaches are more robust, more reusable and can generally avoid the structure-preserving translation of transfer-based approaches to MT. In a corpus-based system, candidate translations can be output with an associated weight or probability and this can facilitate the pruning of a potentially large set of possible translations by a user of the system. New translations can potentially help the system to 'learn' rather than adversely affecting its performance.

Both SMT and EBMT are corpus-based. EBMT systems generally require less training data and, therefore, they are more portable to other language pairs and alternative domains. EBMT systems have traditionally extended correspondences beyond the word-level and most EBMT systems integrate some level of syntactic information. Only recent approaches to SMT have realised the benefits of such measures. Furthermore, the derivation of translation knowledge in an SMT system can sometimes involve elaborate computation, which in an EBMT system is often unnecessary.

The benefits of corpus-based approaches are widely recognised. Despite the benefits of EBMT however, SMT remains the prevalent model for corpus-based research in the field of MT. Furthermore, research classified as EBMT can often involve the integration of rule-based and statistical techniques.

The integration of various techniques in a hybrid environment may ultimately provide the optimal solution to MT. However, if such an advance is ever to be realised, the benefits of individual approaches should be fully appreciated. Moreover, there is much scope for an approach which does not require extensive linguistic resources and can produce reasonably accurate and intelligible translations.

This thesis furthers research in the area of EBMT and explores the application of the Marker Hypothesis in an ⟨English, French⟩ example-based system. We have investigated different dimensions of marker-based EBMT and experimented with applying our *linguistics-lite* methodology in numerous environments.

This work presented in this thesis has:

- shown that an EBMT system which can be developed reasonably quickly using low-level linguistic techniques is useful and extensible to different corpora and domains;
- demonstrated that the Marker Hypothesis can be applied in an ⟨English, French⟩ EBMT system to successfully deduce a set of sub-sententially aligned chunks, words and generalised templates. These can be used to seed the memories of the system and produce novel translations;
- described the implementation of a sub-sentential alignment algorithm and its application to different bitexts;
- highlighted the benefits of the phrasal-lexicon as a means of extending TM technology towards EBMT;
- shown that using on-line MT systems to seed the memories of an EBMT system can generate reasonable results when the source text consists of smaller phrases or is controlled;
- shown that assigning weights to translations based on frequency of occurrence can facilitate the pruning of translation candidates and places the highest-quality translation output by the system in the top 1% of candidates;

- contributed to the integration of the WWW in MT applications by demonstrating that it can be used as a large corpus to validate and correct problems of boundary friction regarding nouns and determiners and nouns and verbs;
- provided the first evidence in favour of performing controlled EBMT and controlling the source language in an EBMT system;
- presented the largest marker-based EBMT system and shown that increasing the training data in a marker-based EBMT system improves translation performance;
- shown that EBMT can outperform SMT and RBMT.

6.1 Contributions of this Thesis

6.1.1 Phrase-Based EBMT

(Carl et al., 2002; Schäler et al., 2003) identified the under-exploited potential of current TM technology. They proposed the transformation of TM to EBMT via the phrasal lexicon. In this thesis, we created a phrase-based EBMT system (Gough et al., 2002; Way and Gough, 2003). We used three on-line rule-based MT systems to derive French translations from 218,697 English phrases extracted from the Penn-II Treebank. Although translations produced via on-line MT systems are generally perceived to be of poor quality, we showed that by seeding an EBMT system with these strings and applying a naïve algorithm to derive a set of smaller correspondences and generalised templates, we could produce translations of high quality. We showed that by combining resources from different on-line systems and increasing our lexical resources we could improve the performance of our system. In a manual evaluation, over 96% of our translations were deemed intelligible. We ranked our translations and showed that the best translation could consistently be located in the top 1% of translations, thus facilitating the pruning of translation candidates presented to a potential user of the system. A manual evaluation showed that we could outperform the RBMT systems used to seed our example-base by up to 50%.

Although not the main objective of our research, as a result of our experiments we were able to evaluate the performance of the individual on-line systems. We found that our results pointed towards *Logomedia* as the best system when compared against *SDL* and *Reverso.*

EBMT systems commonly suffer from problems of boundary friction. We showed that the WWW can be used as a large corpus from which to validate translations produced via EBMT and correct problems of determiner-noun and noun-verb boundary friction. In a *post hoc* stage we improved 82.5% of cases relating to determiner-noun agreement and 76% of cases where noun-verb agreement errors were identified using this method.

6.1.2 Controlled EBMT

In a second experiment, we presented the first controlled EBMT system. Given the unavailability of a suitable bitext, we translated a set of controlled sentences from English-French using the on-line system *Logomedia. Logomedia* was selected for this purpose as it was deemed to be the best of the three on-line systems used to seed the example-base in our phrase-based system (cf. 3.4.6). We applied the Marker Hypothesis to this bitext in order to derive a set of additional chunks, words and generalised templates. We tested our system by translating a set of uncontrolled English and French sentences derived from a *Sun Microsystems* TM. In this way, we reported on a novel experiment where we performed controlled analysis and generation in an EBMT system.

After a number of minimal adjustments to our system, we showed that EBMT could outperform an RBMT system (*Logomedia*) on the same test data and therefore provided support in favour of the hypothesis of (Carl, 2003a; Schäler et al., 2003) that EBMT is possibly more suited to controlled translation than RBMT.

Our results highlighted some anomaly within the automatic evaluation metrics. We provided some discussion and insight into these observations and concluded that in a controlled EBMT system it may be more profitable to exert control over the source data.

We also demonstrated how the integration of lexical correspondence via a dictionary created using the on-line system *Logomedia* could improve the performance of our subsentential alignment algorithm, and consequently improve the quality of our translations by up to 44%.

6.1.3 Scalable EBMT

We scaled up the training data in our EBMT system to include 203,529 ⟨English, French⟩ sentences and consequently we developed the largest English-French EBMT system and the most scalable marker-based system to date. By extracting our sentences from a *Sun Microsystems* TM, we improved the quality of our training data from previous experiments and reduced our dependence on *Logomedia* as a means of seeding our system's resources. We extracted our test set from a subset of the TM, thus increasing the similarity between our test set and our training data under the usual assumption that translating in a similar domain is beneficial to the performance of an EBMT system.

We showed that these factors — scaling up our training data, reducing our dependency on *Logomedia* and increasing the similarity between the training and test data — brought about an improvement in the quality of our translations.

We implemented a novel, length-based filtering technique and showed that this also improved the quality of our translations. We obtained MI scores for the source and target words in our corpus and showed that including MI improved our sub-sentential alignment algorithm, improved and increased the contents of our word-level lexicon and ultimately improved the quality of our translations.

In an automatic evaluation, we compared the translations produced by our system with those derived from an SMT system which integrated word correspondences. Accordingly, we showed that an EBMT system which integrates low-level linguistic techniques can outperform an SMT system which integrates word correspondences.

6.2 Future Research Avenues

6.2.1 Alternative Bitexts

In our phrase-based system (cf. chapter 3), we used three on-line MT systems to translate 218,697 English phrases into French. These systems were chosen on the basis of their ability to process large quantities of text and were not selected on the quality of translations produced. This experiment could be extended by combining resources derived from alternative on-line MT systems. Dummy subjects could also be applied to derive verb forms other than third person singular and plural.

Given that EBMT systems derive translations from a set of real examples, they are perhaps more suited to sublanguage translation and optimal performance can be expected in domain-specific applications. In our scalable system, we carried out EBMT in the domain of computer manual documentation derived from a *Sun Microsystems* TM. TMs are a useful resource and when they are domain-specific are ideal as a bitext for seeding an EBMT system. It would be interesting to extend the Marker Hypothesis to alternative domains and additional text styles.

6.2.2 Other Language Pairs

In this thesis, we applied the Marker Hypothesis to different bitexts. We performed numerous experiments to test the application of our methodology in various contexts.

In our ⟨English, French⟩ EBMT system, we found the Marker Hypothesis to be a useful and efficient method of generating a set of sub-sententially aligned chunks, words and generalised templates. Although the Marker Hypothesis had been applied previously to ⟨English, French⟩ in the METLA system (Juola, 1994, 1997), the experiment was on an extremely small scale and the example-base consisted of just 29 sentence pairs. English and French have similar word order and are both SVO languages. Nevertheless, translation from English to French does involve cases of complex transfer and can incur problems of boundary friction. However, one question which now presents itself is — how applicable is the Marker Hypothesis to other language pairs?

The Marker Hypothesis has been applied, albeit on a small scale, to ⟨English,Urdu⟩ which are typologically different languages. It has also been applied to ⟨English, German⟩ in the *Gaijin* system. It is said to be a universal constraint and has been shown to facilitate languages with and without specific marker words (Green, 1979). Furthermore, according to (Juola, 1994) marker constructs do exist universally in all languages. As such, there would appear to be great scope for extending the Marker Hypothesis to other language pairs.

Nevertheless, a number issues of would need to be addressed. A bitext would be required for the languages in question. Given that we have shown that a more scalable training data set can improve translation performance and overcome problems of data-sparseness, this corpus should be sizeable. A larger corpus will also add reliability to translation weights and MI scores. Likewise, a set of marker words for each language would need to be obtained. This would require some human expertise for both languages. The sub-sentential alignment algorithm was implemented with the intention of making it as portable as possible to other language pairs. A dictionary such as that obtained in our system from *Logomedia* could be added for a new language pair. MI can be calculated from the bitext and does not require additional resources.

Some initial experiments have been performed in an effort to extend marker-based EBMT to other language pairs. For example, work has been initiated for English-Chinese translation with colleagues from Harbin Institute of Technology in China. Initial results, although unpublished, appear to be encouraging. The Chinese language does not make as much use of the determiner class as English. However, marker categories can be added and removed liberally to adapt to the language pairs involved. For example, for Chinese-English, we have included punctuation as a marker class <PNCT>.

We have also investigated applying the Marker Hypothesis to English-Irish translation. Unlike English and French, Irish is a Verb-Subject-Object (VSO) language. However, we can successfully apply the Marker Hypothesis to produce useful sub-sentential alignments. For example, consider the sentence-aligned pair in (174):

(174) “ Chuaigh an fear go dtí an siopa. ”
“ went the man to the shop ”
“ The man went to the shop.”

The past tense verb *went* translates as the verb *chuaigh* which appears at the beginning of the Irish sentence. The preposition *to* in English also translates as two words in Irish *go* and *dtí.* If we apply the Marker Hypothesis to the sentence pair in (174), we obtain the fragmented sentences in (175):

(175) <DET> the man went <PREP> to <DET> the shop ⇔ <NULL> Chuaigh <DET> an fear <PREP> go <PREP> dtí <DET> an siopa

Assuming that a dictionary or MI exists which can link the words in (176), we can establish the 1:2 and 1:1 alignments in (177).

(176) man ⇔ fear
went ⇔ chuaigh
shop ⇔ siopa

(177) <DET> the man went ⇔ chuaigh an fear
<PREP> to the shop ⇔ go dtí an siopa

The sub-sentential alignments in (177) could subsequently be generalised by replacing the marker words with their tags to form the templates in (178):

(178) <DET> man went ⇔ chuaigh <DET> fear
<PREP> <DET> shop ⇔ <PREP> <PREP> <DET> siopa

Experiments with English-Irish marker-based EBMT are ongoing but initial results suggest that such an approach is productive.

6.2.3 Recombination

We replace marker words with their associated tag and in so doing make our matching algorithm more flexible. This also facilitates the recombination process as only certain word classes can appear in a specified context, similar to the 'hooks' applied in (Somers et al., 1994). We also correct some instances of determiner-noun and noun-verb boundary friction *post hoc*.

Given that English and French are SVO languages we use this information in conjunction with our generalised templates to order the chunks retrieved from our lexicons. We find that our chunks retain enough context so that this does not present us with major problems. However, this method is not extensible to every language pair. Several techniques could be applied to enhance the recombination process. The *post hoc* validation process could be extended to reorder words if one sequence receives a higher number of hits than its alternatives. Another possibility might be to use a method similar to (Somers et al., 1994) or McTait (2003) (cf. section 2.4.6) where the corpus itself is used to check that strings are well-formed.

Without integrating any further linguistic resources, we could develop the current templates further. For example, we could extend our marker categories to include a class <NUM> which marks numbers. (Brown, 1999) (cf. section 2.4.4) identifies certain classes of words such as dates, place names etc., which he terms 'placeables'. Using a similar method, we could replace numbers with a general tag. For example, the aligned pair in (179) could be generalised to form the generalised pair in (180):

(179) 10 green bottles ⇔ 10 bouteilles vertes

(180) <NUM> green bottles ⇔ <NUM> bouteilles vertes

This would allow an input string such as *14 green bottles* to be matched against the source side of the template in (180).

In most cases, the integration of our dictionary allows for word order to be preserved within the context of the sub-sententially aligned chunks. However, we could also use this dictionary to establish links between content words and replace these with relevant tags. For example, assume the chunk pair in (181) occurs in our marker-lexicon:

(181) the blue house ⇔ la maison bleue

Now assume that we want to translate the string *the red house.* If this string or a generalised template <DET> *red string* is not located in our system's memories the string will be translated word for word. While we could integrate methods similar to those described in (Somers et al., 1994; McTait, 2003) to smooth over the word order in the string, we could also avoid the problem by generalising the content words in the chunk in (181). We can use the dictionary to establish a correspondence between *blue* and *bleue* and between *house* and *maison.* We can then generalise the linked content words to form the generalised templates in (182):

(182) a. the blue <LEX> ⇔ la <LEX> bleue

b. the <LEX> house ⇔ la maison <LEX>

If we generalise the string *the red house* in the same way, one template produced *the* <LEX> *house* can be matched against the template in (182b). The corresponding target template *la maison*<LEX> can be retrieved and the translation of the generalised content word *red* can be inserted in the position <LEX> in the target template, assuming, of course, that this word occurs in the word-level lexicon. This would produce the translation in (183):

(183) la maison rouge

6.2.4 Efficiency

The efficiency of our system should also be addressed with any further development. The system is currently implemented in PERL. Although PERL is useful for language processing and sufficient for the experiments outlined in this thesis, training our system on an even more scalable data set or extending the system to interact with a human user such as to improve on current TM technology could prove more productive if the system were to be implemented in a more modular fashion using *C++* or Java.

One approach to improving the efficiency of an example-based system is proposed in (Brown, 2004) which describes the application of the Burrows-Wheeler Transform (BWT) to an EBMT system. The BWT is adapted to word-based indexing of the training corpus and this facilitates the development of a scalable model by matching training instances without requiring additional space. All instances of an n-grams are grouped together and the result is an order-of-magnitude speedup at run time.

6.3 Closing Remarks

We have shown that our EBMT system can produce good results in different environments without recourse to complex linguistics resources or techniques. We have also demonstrated that a number of additional techniques could further improve the performance of the system while maintaining a *linguistics-lite* methodology.

The potential benefits of integrating methods in a hybrid system cannot be overlooked. Inferring 'rules' from corpora may be a way to revive rule-based techniques, while the integration of phrases and syntactic information in SMT may optimise the performance of corpus-based systems. Given the advantages and applicability of EBMT, it is likely that it will play a vital role in the development of any successful hybrid system.

Bibliography

Akis, J. W. and Sisson, W. R. (2002). Improving Translatability: A Case Study at Sun Microsystems, Inc. In *The LISA Newsletter*, volume **11**.

Andriamanankasina, T., Araki, K., and Tochinai, K. (2003). EBMT of POS-Tagged Sentences via Inductive Learning. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 225–252. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Aue, A., Menezes, A., Moore, B., Quirk, C., and Ringger, E. (2004). Statistical Machine Translation Using Labeled Semantic Dependency Graphs. In *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 125–134, Baltimore, MD.

Becker, J. (1975). The Phrasal Lexicon. In *Proceedings of the International Workshop on Theoretical Issues in Natural Language Processing*, pages 70–73, Cambridge, MA.

Berlin, B. and Kay, P. (1969). *Basic Color Terms : Their Universality and Evolution.* University of California Press, Berkeley, CA.

Bernth, A. (2003). Controlled Generation for Speech-to-Speech MT Systems. In *Proceedings of Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation (EAMT-CLAW-03)*, pages 1–7, Dublin, Ireland.

Block, H.-U. (2000). Example-Based Incremental Synchronous Interpretation. In Wahlster, W., editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 411–417. Springer, Heidelberg, Germany.

Bond, F. and Shirai, S. (2003). A Hybrid Rule and Example-Based Method for Machine Translation. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 211–224. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Brown, P., Cocke, J., Pietra, S. D., Jelinek, F., Mercer, R., and Roossin, P. (1988). A Statistical Approach to Language Translation. In *Proceedings of the 12th International Conference on Computational Linguistics, (COLING-88)*, pages 71–76, Budapest, Hungary.

Brown, R. (2004). A modified burrows-wheeler transform for highly scalable example-based translation. In Frederking, R. E. and Taylor, K. B., editors, *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-04)*, pages 27–36, Georgetown, Washington DC.

Brown, R. D. (1999). Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32, Chester, UK.

Brown, R. D. (2002). Automated Generalization of Translation Examples. In *Proceedings of the 18th International Conference on Computational Linguistics (Coling-02)*, pages 672–678, Saarbrücken, Germany.

Brown, R. D. (2003). Clustered Transfer Rule Induction for Example-Based Translation. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 287–307. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Carl, M. (2003a). Data-Assisted Controlled Translation. In *Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation (EAMT-CLAW 03)*, pages 16–24, Dublin, Ireland.

Carl, M. (2003b). Inducing Translation Grammars from Bracketed Alignments. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 339–361. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Carl, M., Way, A., and Schäler, R. (2002). Toward a Hybrid Integrated Translation Environment. In Richardson, S., editor, *Machine Translation: From Research to Real Users: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 11–20, Tiburon, CA.

Chandioux, J. (1976). MÉTÉO: un système opérationnel pour la traduction automatique des bulletins météreologiques destinés au grand public. *META*, **21**:127–133.

Charniak, E., Knight, K., and Yamada, K. (2003). Syntax-based language models for statistical machine translation. In *Proceedings of MT Summit IX*, pages 40–46, New Orleans, LA.

Chomsky, N. (1981). *Lectures on Government and Binding.* Foris Publications, Dordrecht, The Netherlands.

Church, K. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, **16**(1):22–29.

Cicekli, I. and Güvenir, H. A. (2003). Learning Translation Templates from Bilingual Translation Examples. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 255–287. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Collins, B. (1998). *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach.* PhD thesis, Trinity College, Dublin.

Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. In *Proceedings of MT Summit IX*, pages 63–70, New Orleans, LA.

Cranias, L., Papageorgiou, H., and Piperidis, S. (1994). A Matching Technique in Example-Based Machine Translation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-94)*, pages 100–104, Kyoto, Japan.

Cranias, L., Papageorgiou, H., and Piperidis, S. (1997). Example Retrieval from a Translation Memory. *Natural Language Engineering*, **3**(4):255–277.

Dabbadie, M., Hartley, A., King, M., Miller, K. J., Hadi, W. M. E., Popescu-Belis, A., Reeder, F., and Vanni, M. (2002). A hands-on study of the reliability and coherence of evaluation metrics. In *Proceedings of the Workshop at the 3rd International Conference on Language Resources and Evaluation (LREC-02)*, pages 8–16, Las Palmas, Canary Islands, Spain.

Frederking, R. and Brown, R. (1996). The Pangloss-Lite Machine Translation System. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas*, pages 268–272, Montreal, Canada.

Furuse, O. and Iida, H. (1992a). An Example-Based Method for Transfer-Driven Machine Translation. In *Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, pages 139–150, Montreal, Canada.

Furuse, O. and Iida, H. (1992b). Cooperation between Transfer and Analysis in Example-Based Framework. In *Proceedings of 15th [sic] International Conference on Computational Linguistics (COLING-92)*, volume **2**, pages 645–651, Nantes, France.

Gough, N. and Way, A. (2003). Controlled Generation in Example-Based Machine Translation. In *Proceedings of MT Summit IX*, pages 133–140, New Orleans, LA.

Gough, N. and Way, A. (2004). Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pages 95–104, Baltimore, MD.

Gough, N., Way, A., and Hearne, M. (2002). Example-Based Machine Translation via the Web. In Richardson, S., editor, *Machine Translation: From Research to Real Users: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 74–83, Tiburon, CA. Springer, Heidelberg, Germany.

Green, T. R. (1979). The Necessity of Syntax Markers: Two Experiments with Artificial Languages. *Verbal Learning and Verbal Behavior*, **18**:481–496.

Greenberg, J. H. (1966). *Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements.* MIT Press, Cambridge, MA.

Grefenstette, G. (1999). The World Wide Web as a Resource for Example-Based Machine Translation. In *Proceedings of the ASLIB Conference on Translating and the Computer*, number 21 [pages not numbered], ASLIB/IMI, London.

Hain, T., Woodland, P., Niesler, T., and Whittacker, E. (1998). The 1998 htk system for transcription of conversational telephone speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP-99)*, volume **1**, pages 57–60, Phoenix, AZ.

Hartley, A., Scott, D., Bateman, J., and Dochev, D. (2001). AGILE- A System for Multilingual Generation of Technical Instructions. In Maegaard, B., editor, *Proceedings of MT Summit VIII, Machine Translation in the Information Age*, pages 145–150, Santiago de Compostela, Spain.

Hearne, M. and Way, A. (2003). Seeing the Wood for the Trees: Data-Oriented Translation. In *Proceedings of MT Summit IX*, pages 165–172, New Orleans, LA.

Hein, A. S. (1996). Preference Mechanisms of the Multra Machine Translation System. In *Discourse and Meaning. Papers in Honour of Eva Hajicova*, pages 321–333, Amsterdam and Philadelphia. John Benjamins Publishing Company.

Hovy, E. (1998). Generating Language with a Phrasal Lexicon. In *Natural Language Generation Systems*, pages 353–384. Springer Verlag, New York, NY.

Hutchins, J. and Somers, H. (1992). *An Introduction to Machine Translation.* Academic Press, London.

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

Juola, P. (1997). Corpus-based Acquisition of Transfer Functions using Psycholinguistic Principles. In Jones, D. and Somers, H., editors, *New Methods in Language Processing*, pages 207–218. UCL Press, London.

Juola, P. (1998). On Psycholinguistic Grammars. *Grammars*, **1**(1):15–31.

Jurafsky, D. and Martin, J. H. (2002). *Speech and Language Processing: An Introduction to Natural Language Procesing, Computational Linguistics and Speech Recognition.* Prentice Hall, Upper Saddle River, NJ.

Kaji, H., Kida, Y., and Morimoto, Y. (1992). Learning Translation Templates from Bilingual Text. In *Proceedings of the 15th [sic] International Conference on Computational Linguistics (COLING-92)*, pages 672–678, Nantes, France.

Kamprath, C., Adolphson, E., Mitamura, T., and Nyberg, E. (1998). Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English. In *Proceedings of the 2nd International Workshop on Controlled Language Applications (CLAW 98)*, pages 51–61, Pittsburgh, PA.

Keenan, E. and Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, **8**:63–99.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Human Language Technology Conference, (HLT-NAACL)*, pages 48–54, Edmonton, Canada.

Kulesza, A. and Shieber, S. M. (2004). A learning approach to improving sentence-level mt evaluation. In *Proceedings of the 10th Conference on Theoretical and Methodological Issues in Machine Translaiton (TMI-04)*, pages 75–84, Baltimore, MD.

Langlais, P. and Simard, M. (2002). Merging Example-Based and Statistical Machine Translation: An Experiment. In *Machine Translation: From Research to Real Users: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 104–113.

Levenshtein, V. (1965). Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones. In *Problems of Information Transmission*, number 1, pages 8–17.

Littlestone, N. and Warmuth, M. (1992). The Weighted Majority Algorithm. Technical Report USCS-CRL 91.28, University of California, Santa Cruz, CA.

Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP-2002)*, pages 133–139, University of Pennsylvania, Philadelphia, PA.

McTait, K. (2003). Translation Patterns, Linguistic Knowledge and Complexity in EBMT. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 307–338. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Means, L. and Godden, K. (1996). The Controlled Automotive Service Language (CASL) Project. In *Proceedings of the 1st International Workshop on Controlled Language Applications (CLAW-96)*, pages 106–114, Leuven, Belgium.

Melamed, D., Green, R., and Turian, J. (2003). Precision and recall of machine translation. Technical Report 03-004, New York University, NY.

Menezes, A. and Richardson, S. D. (2003). A Best-First Alignment Algorithm for Extraction of Transfer Mappings. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 421–442. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Milosavljevic, M., Tulloch, A., and Dale, R. (1996). Text Generation in a Dynamic Hypertext Environment. In *Proceedings of the 19th Australiasian Computer Science Conference*, pages 417–426, Melbourne, Australia.

Mima, H., Iida, H., and Furuse, O. (1998). Simultaneous Interpretation Utilizing Example-based Incremental Transfer. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 855–861, Montreal, Canada.

Mitamura, T. and Nyberg, E. (1995). Controlled English for Knowledge Based MT: Experience with the KANT System. In *Proceedings of the 6th International Conference on Theoretical and Methodlogical Issues in Machine Translation (TMI-95)*, pages 158–172, Leuven, Belgium.

Morgan, J., Meier, R. P., and Newport, E. L. (1989). Facilitating the Acquisition of Syntax with Cross-Sentential Cues to Phrase Structure. *Journal of Memory and Language*, **28**:360–374.

Mori, K. and Moeser, S. D. (1983). The Role of Syntax Markers and Semantic Referents in Learning an Artificial Language. *Journal of Verbal Learning and Verbal Behavior*, **22**:701–718.

Murata, M., Ma, Q., Uchimoto, K., and Isahara, H. (1999). An Example-Based Approach to Japanese-to-English Translation of Tense, Aspect and Modality. In *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 66–76, Chester, UK.

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In Elithorn, A. and Banerji, R., editors, *Artificial and Human Intelligence*, pages 173–180. Amsterdam, The Netherlands.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(2):19–51.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD.

Öz, Z. and Cicekli, I. (1998). Ordering Translation Templates by Assigning Confidence Factors. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA-98)*, pages 51–61, Langhorne, PA.

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

Planas, E. and Furuse, O. (2003). Formalizing Translation Memory. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 157–188. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Poutsma, A. (2003). Machine Translation with Tree-Dop. In Bod, R., Scha, R., and Simaán, K., editors, *Data-Oriented Parsing*, pages 339–357.

Power, R., Scott, D., and Hartley, A. (2003). Multilingual Generation of Controlled Languages. In *Proceedings of Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Controlled Translation (EAMT-CLAW 03)*, pages 115–123, Dublin, Ireland.

Rayner, M. and Carter, D. (1997). Hybrid Language Processing in the Spoken Language Translator. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 107–110, Munich, Germany.

Sato, S. (1993). Example-Based Translation of Technical Terms. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 58–68, Kyoto, Japan.

Sato, S. (1995). MBT2: A Method for Combining Fragments of Examples in Example-Based Machine Translation. *Artificial Intelligence*, **75**:31.

Sato, S. and Nagao, M. (1990). Toward Memory-Based Translation. In *Proceedings of the 13th Conference on Computational Linguistics (COLING-90)*, pages 247–252, Helsinki, Finland.

Schäler, R. (1996). Machine translation, translation memories and the phrasal lexicon: The localisation perspective. In *Proceedings of the 4th International Congress on Terminology and Knowledge Engineering (TKE-96), EAMT Workshop on Machine Translation*, pages 21–33, Vienna, Austria.

Schäler, R., Way, A., and Carl, M. (2003). Example-Based Machine Translation in a Controlled Environment. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 83–114. Kluwer Academic Publishers, Dordrecht, The Netherlands.

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

Somers, H. (2003). An Overview of EBMT. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 3–57. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Somers, H., McLean, I., and Jones, D. (1994). Experiments in multilingual example-based generation. In *3rd International Conference on the Cognitive Science of Natural Language Processing (CSNLP-94)*, [pages not numbered], Dublin, Ireland.

Soricut, R., Knight, K., and Marcu, D. (2002). Using a large monolingual corpus to improve translation accuracy. In Richardson, S., editor, *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA-2002)*, pages 155–164, Tiburon, CA. Springer, Heidelberg, Germany.

Sumita, E. (2003). EBMT Using DP-matching Between Word Sequences. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 189–209. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Sumita, E. and Iida, H. (1991). Experiments and Prospects of Example-Based Machine Translation. In *29th Annual Meeting of the Association for Computational Linguistics (ACL-91)*, pages 185–192, Berkeley, CA.

Sumita, E., Iida, H., and Kohyama, H. (1990). Translating with Examples: A new Approach to Machine Translation. In *The 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language (TMI-90)*, pages 203–212, Austin, TX.

Talmy, L. (1988). The relation of grammar to cognition. In Rudzka-Oysten, B., editor, *Topics in Cognitive Linguistics*. John Benjamins Publishing Co., Amsterdam and Philadelphia.

Turcato, D. and Popowich, F. (2003). What is Example-Based Machine Translation. In *Recent Advances in Example-Based Machine Translation*, pages 59–81. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Turian, J., Shen, L., and Melamed, D. (2003). Evaluation of Machine Translation and its Evaluation. In *Proceedings of MT Summit IX*, pages 386–393, New Orleans, LA.

van der Eijk, P., de Koning, M., and van der Steen, G. (1996). Controlled Language Correction and Translation. In *Proceedings of the 1st International Workshop on Controlled Language Applications (CLAW 96)*, pages 64–73, Leuven, Belgium.

van Rijsbergen, C. J. (1979). *Information Retrieval.* Butterworth, London.

Veale, T. and Way, A. (1997). Gaijin: A Bootstrapping Approach to Example-Based Machine Translation. In *International Conference, Recent Advances in Natural Langugage Processing*, pages 239–244, Tzigov Chark, Bulgaria.

Wagner, R. A. and Fischer, M. J. (1974). The String-to-String Correction Problem. *Journal of the Association for Computing Machinery*, **18**:168–173.

Wang, Y. (1998). *Grammar Inference and Statistical Machine Translation.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

Watanabe, H. (1992). A Similarity-Driven Transfer System. In *Proceedings of the 15th [sic] International Conference on Computational Linguistics (COLING-92)*, pages 770–776, Nantes, France.

Watanabe, H., Kuroahashi, S., and Aramaki, E. (2003). Finding Translation Patterns from Dependency Structures. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 397–420. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Watanabe, H., Kurohashi, S., and Aramak, E. (2000). Finding Structural Correspondences from Bilingual Parsed Corpora for Corpus-based Translation. In *Proceedings of the 18th International Conference on Computatonal Linguistics (Coling-00)*, Saarbrücken, Germany.

Way, A. (2001). *LFG-DOT: A Hybrid Architecture for Robust MT.* PhD thesis, Department of Language and Linguistics, University of Essex, Colchester, UK.

Way, A. (2003). Translating with Examples: The LFG-DOT Models of Translation. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 443–472. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Way, A. and Gough, N. (2003). Example-Based Machine Translation using the World Wide Web. *Computational Linguistics*, **29**(3).

Way, A. and Gough, N. (2004). Example based Controlled Translation. In *Proceedings of the 9th Workshop of the European Association for Machine Translation (EAMT)*, pages 73–81, Valetta, Malta.

Way, A. and Gough, N. (forthcoming). Comparing a Scalable Marker-Based EBMT System with SMT. *Natural Language Engineering.*

Xia, F. and McCord, M. (2004). Improving a Statistical MT system with Automatically Learned Rewrite Patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (Coling-04)*, pages 508–514, Geneva, Switzerland.

Yamada, K. and Knight, K. (2001). A Syntax-Based Statistical Translation Model. In *29th Annual Meeting of the Association for Computational Linguistics (COLING-01)*, pages 523–530, Toulouse, France.

Yamamoto, K. and Matsumoto, Y. (2003). Extracting Translation Knowledge from Parallel Corpora. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 365–395. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Zernik, U. and Dyer, M. (1997). The Self-Extending Phrasal Lexicon. *Computational Linguistics*, **13**(3-4):308–327.