Corpus-based machine translation evaluation via automated error detection in output texts

Debra Elliott

Submitted in accordance with the requirements for
the degree of Doctor of Philosophy

The University of Leeds
School of Computing and
Centre for Translation Studies

September, 2006

Acknowledgements

First and foremost, I would like to thank my husband John for tolerating my ups and downs during the last four years and for finding other things to do while I have been chained to my laptop. Thankfully, he understood exactly what I was going through, as he completed his PhD only three years ago.

Thank you also to my two supervisors - Eric Atwell in the School of Computing and Professor Anthony Hartley in the Centre for Translation Studies - for their time, encouragement and constructive feedback.

Finally I would like to thank fellow PhD student, Andy Roberts, for helping me to learn computer programming from scratch and to Dr Serge Sharoff, Lecturer in the Centre for Translation Studies, for his help in using the BootCat Toolkit.

Abstract

Since the emergence of the first fully automatic machine translation (MT) systems over fifty years ago, the use of MT has increased dramatically. Consequently, the evaluation of MT systems is crucial for all stakeholders. However, the human evaluation of MT output is expensive and time-consuming, often relying on subjective quality judgements and requiring human 'reference translations' against which the output is compared. As a result, interest in more recent years has turned towards automated evaluation methods, which aim to produce scores that reflect human quality judgements.

As the majority of published automated evaluation methods still require human 'reference translations' for comparison, the goal of this research is to investigate the potential of a method that requires access only to the translation. Based on detailed corpus analyses, the primary aim is to devise methods for the automated detection of particular error types in French-English MT output from competing systems and to explore correlations between automated error counts and human judgements of a translation as a whole.

First, a French-English corpus designed specifically for MT evaluation was compiled. A sample of MT output from the corpus was then evaluated by humans to provide judgements against which automated scores would ultimately be compared. A data-driven fluency error classification scheme was subsequently developed to enable the consistent manual annotation of errors found in the English MT output, without access to the original French text. These annotations were then used to guide the selection of error categories for automated error detection, and to facilitate the analysis of particular error types in context so that appropriate methods could be devised. Manual annotations were further used to evaluate the accuracy of each automated approach. Finally, error detection algorithms were tested on English MT output from German, Italian and Spanish to determine the extent to which methods would need to be adapted for use with other language pairs.

Contents

List of Figures

List of Tables

Acronyms and abbreviations

AAMT	Asia-Pacific Association for Machine Translation
ALPAC	Automatic Language Processing Advisory Committee
AMTA	Association for Machine Translation in the Americas
ATA	American Translators' Association
BLEU	BiLingual Evaluation Understudy
BNC	British National Corpus
CAT	Computer-Assisted / Computer-Aided Translation
DARPA	Defense Advanced Research Projects Agency
DiET	Diagnostic and Evaluation Tools
EAMT	European Association for Machine Translation
EBMT	Example-Based Machine Translation
EC	European Commission
ECMT	European Commission Machine Translation
EU	European Union
FAQs	Frequently Asked Questions
FEMTI	Framework for the Evaluation of Machine Translation in ISLE
FOLDOC	Free On-Line Dictionary of Computing
FSF	Free Software Foundation
GNU	Gnu Not UNIX
GTM	General Text Matcher
HLT	Human Language Technology
HT	Human translation
IAMT	International Association for Machine Translation
ISLE	International Standards in Language Engineering
ITR	International Translation Resources
JEIDA	Japanese Electronic Industry Development Association
KBMT	Knowledge-Based Machine Translation
LTV	Legitimate Translation Variation
MT	Machine translation
NE	Named entity
NEVA	N-Gram Evaluation
NIST	National Institute of Standards and Technology

NLP	Natural Language Processing
NLTK	Natural Language ToolKit
NP	Noun phrase
PoS	Part(s) of speech
RED	Ranker based on Edit Distances
RST	Rhetorical Structure Theory
SAE	Society of Automotive Engineers
SER	Sentence Error Rate
SMT	Statistical Machine Translation
SL	Source language
SSER	Subjective Sentence Error Rate
ST	Source text
SVM	Support vector machine
TAPs	Think Aloud Protocols
TIDES	Translingual Information Detection, Extraction and Summarization
TL	Target language
TM	Translation Memory
TSNLP	Test Suites for Natural Language Processing
TT	Target text
WER	Word Error Rate

Chapter 1

Introduction: problems and objectives

« La traduction automatique ne fonctionne pas parfaitement. Et heureusement, car elle est une source de réflexions infinies. »

Anne-Marie Loffler-Laurian

"Machine translation does not function perfectly. And fortunately, because it is a source of infinite reflexions."

Translation by Systran Version 4.0

Since the emergence of the first fully automatic machine translation (MT) systems over fifty years ago, the use of MT to translate texts from one natural language into another, in a variety of contexts and for a number of different purposes, has increased dramatically. Globalisation and the availability of texts in a multitude of languages on the Internet have led to a growing need for quick and cheap translations of varying levels of quality depending on user requirements. As a result, many companies and government organisations throughout the world are using tailor-made or off-the-shelf MT systems, and the number of users of free online MT is now greater than ever (see section 2.2). The evaluation of MT systems, both in terms of their performance as Natural Language Processing (NLP) applications, and the quality of the translations they produce, is therefore crucial for investors, researchers, developers, vendors, project managers and end-users.

The evaluation by humans of machine translated texts, whether to compare the quality of output from competing systems or to monitor improvements in output from the same system during development, is a complex, expensive and time-consuming task. Unlike the evaluation of part-of-speech taggers, parsers or speech recognisers (Atwell et al., 2000) it is not simply a matter of comparing MT output to some "gold standard" human translation, since translation is legitimately subject to stylistic and other variation. Instead, MT evaluation relies on either the objective scoring of very specific linguistic

phenomena using test suites, or the somewhat subjective quality judgements made by evaluators who are trained to score segments (often sentences) of translated text using a particular metric. For example, evaluators might rate fluency (the extent to which the translation reads like natural English written by a native speaker) or fidelity ("the degree to which the information contained in the original text has been reproduced without distortion in the translation" (Van Slype, 1979)) using a scoring scale to rate each segment.

MT evaluation is expensive for a number of reasons: in order to reduce the problem of subjectivity, scores must be obtained from several evaluators and, to improve the reliability of results, a large number of texts should be used. Furthermore, for the evaluation of fidelity, bilingual evaluators are required to compare the MT output against the source text (ST) or alternatively, human "reference translations" must be produced, against which evaluators with no knowledge of the source language (SL) can compare and score the MT output. The employment of several judges to evaluate a large number of texts, along with the possible requirement of bilingual knowledge or human translations, makes MT evaluation very costly and time-consuming.

During the last five decades, a vast number of different evaluation methods have been explored and tested, all making their contributions to progress in the field and leading, in more recent years, to an interest in quicker and cheaper automated methods, which aim to produce scores that correlate highly with human judgements. A detailed critical analysis of human and automated approaches is presented in Chapter 3.

Current automated methods, however, have their drawbacks. For instance, the widely-used BLEU (BiLingual Evaluation Understudy) method (Papineni et al., 2001) requires the expense of producing up to four human reference translations of each source text, against which the machine translations are automatically compared and scored according to modified *n*-gram precision. The RED (Ranker based on Edit Distances) approach (Akiba et al., 2001) also uses multiple reference translations to automatically rank MT output based on edit distances. In one published evaluation (Akiba et al., 2003), 16 human translations of 345 sentences in two language directions were used from the Basic Travel Expression Corpus (Takezawa et al., 2002). Each of the above approaches requires human reference translations for every language pair to be

evaluated and these tend to be reused in subsequent evaluations to save time and expense, even perhaps in cases where different kinds of texts would be more appropriate for the needs of the stakeholders in question. A further drawback is that they do not help system developers or MT post-editors by highlighting individual errors in the machine translations, nor is there a straightforward correlation with human judgements.

In response to these drawbacks, the purpose of this research is to explore the potential of a fully automated evaluation method that does not involve human translations; does not require access to, or the ability to understand, the source text. The potential of such a method, which evaluates a text based on the analysis of the MT output alone, is supported by the fact that in many human evaluations, scores representing fidelity correlate highly with judgements of fluency, where scores are based on a reading of the MT output alone (eg. Carroll, 1966; White et al., 1994; White, 2001; Elliott et al., 2004a).

1.1 Objectives

The primary aim was to investigate the potential of a fully automated method for the detection of errors in MT output, which produces scores that correlate with human quality judgements and which does not require access to human translations or to the source text. This would provide the basis for creating an intelligent system that can detect and annotate different categories of errors that typically occur in MT output, incorporating human language processing rationales developed from initial supervised methods.

Automatically detected errors would then be analysed to establish whether the number of errors found in particular categories would enable the generation of a score that correlates highly with human judgements. Initial experiments would focus on the evaluation of texts in the technical domain, reflecting the needs of many MT users (see 4.2.1), translated from French into English. The proposed method has a number of advantages over existing approaches:

- It eliminates the expense of producing human translations and enables users to evaluate any text sample without relying on pre-translated source texts;

- Automatic error detection would be of use to MT post-editors, who need to quickly identify errors for revision, and would help developers to pinpoint areas for improvement;
- It should be easily extendible to other language pairs in which the target language (TL) is English;
- Unlike methods such as BLEU (Papineni et al., 2001), the new approach should ultimately be able to take the gravity of each error type into account by introducing weighted scores, rather than assuming that all tokens in a text are of equal importance.

The design of a template for such an automated error detection system involved a series of sub-goals:

- Machine translation systems are used by companies, organisations and individuals throughout the world, and it was important that texts selected for this research reflected real use of MT. Decisions had to be made on the type, length and number of texts to be used. The first sub-goal was, therefore, to design and compile a corpus specifically for this research.

- Any reliable method for the automatic evaluation of MT output must produce scores that reflect human judgements. Having machine-translated texts from the corpus using available systems, the second goal was to obtain human scores for each translation, based on established evaluation methods, against which automated scores would subsequently be compared.

- Machine translations contain a number of different error types, ranging from inappropriate word choices to incorrect syntax. The next objective was to manually identify, classify and annotate all errors in the same sample of MT output, without access to the source text. The purpose of this was to guide decisions on which error categories to select for automatic detection, and to facilitate the analysis of particular error types in context, so that appropriate automated detection methods could be devised.

- Having established a classification of errors, the next goal was to select a number of more frequent and easily detectable error types for automatic identification. For each

selected error category, the aim was to devise a detection method, using human language processing rationales developed from initial supervised approaches. The reliability of each algorithm would then be tested by comparing automatically detected errors in each category with the manual annotations.

- The next objective was to identify a set of errors that, when automatically detected, would generate scores that correlate highly with human judgements. This stage would involve a detailed comparative analysis of human scores and automated error counts.

- The final goal was to determine the extent to which the algorithms would need to be adapted, if at all, for use with other language pairs. The developed methods would be tested on MT output translated from different source languages into English.

1.2 Thesis outline

The chapters of this thesis reflect the order of the objectives outlined above. However, by way of an introduction to this project, it is important first of all to provide some context in the form of an overview of machine translation systems and their use. Chapter 2 begins with a definition of machine translation for the purpose of this research and lists a number of reasons why it is needed. A brief history of MT and its growing use is presented, highlighting the importance of research in MT evaluation. The chapter briefly describes how machine translation engines work, why translation is difficult for computers and, therefore, why particular errors in MT output can occur.

The aim of Chapter 3 is to describe and critically analyse the various different approaches to MT evaluation. First, published work on the human evaluation of MT output is explored. This investigation would guide the selection of one or more human evaluation methods for this research. The second aim of this chapter is to present and analyse existing automated methods for MT evaluation, including BLEU (Papineni et al., 2001) and RED (Akiba et al., 2001). The advantages and disadvantages of each method are investigated, and the complexities involved are discussed before decisions are taken on directions for this research.

Chapter 4, based on (Elliott et al., 2003, 2004a, 2004b), describes the design, compilation and content of a new corpus for MT evaluation. A detailed rationale for

corpus design is provided, including reasons for decisions on corpus size, structure and content. This is based on an investigation of existing corpora for MT evaluation and a survey of MT users. The subsequent human evaluations of a sample of texts from the corpus are described, and results are presented. Finally, the development of a hierarchical MT error classification scheme, using the same sample of texts, is outlined. The scheme and examples of corpus annotation are also presented, along with error statistics for all systems and texts, to guide the development of automated error detection algorithms.

Chapter 5, much of which is based on (Elliott et al., 2005), details the development of algorithms for the automated detection of five selected error types: 'outrageous' words[1], inappropriate content words, inappropriate prepositions, inappropriate pronouns and untranslated words and acronyms. The accuracy of each algorithm is evaluated in turn, based on comparisons between the number of automatically detected errors and manual error annotations. Comparisons are also investigated between automated scores, based on the detection of individual error types, and human judgements, to ascertain whether the automated detection of one error type could predict the quality of a translation as a whole. Automated scores from all possible combinations of error categories are then compared with human scores to determine which combination generates a score that correlates most highly with human quality judgements.

Chapter 6 presents findings obtained when the algorithms developed for the French-English language pair were tested on English output translated from three other source languages: German, Italian and Spanish. Automated and manual annotations are compared to investigate the extent to which the existing methods would need to be adapted to detect a similar proportion of errors in English MT translated from other languages.

Conclusions are presented in Chapter 7. Goals and achievements are summarised and suggestions for future work are discussed.

---

[1] These words are highly unusual in a given text type and domain. In most cases, outrageous words are mistranslations which occur when a source language word has two or more valid translations in the target language, with very different meanings.

Chapter 2

Background

The aim of this chapter is to provide a wider context for the current research by first presenting an overview of the development and use of MT and second, describing MT systems and their architectures, in order to gain an understanding of the quality of the translations they produce. By way of an introduction to this, Section 2.1 begins with a definition of machine translation for the purpose of this research, and provides several reasons why MT is needed. This is followed by a brief history of MT and its use in 2.2, highlighting the growing need world-wide for quick and cheap translations and hence the importance of MT evaluation for system development and comparison. Section 2.3 describes how MT systems work, with a concise look at MT architectures by way of an introduction to 2.4, which describes why translation is difficult for computers and, therefore, why particular kinds of errors are made. Conclusions drawn from this background chapter are presented in 2.5.

2.1 What is machine translation and why do we need it?

According to (Hutchins and Somers, 1992), "the term Machine Translation (MT) is the now traditional and standard name for computerized systems responsible for the production of translations from one natural language to another, with or without human assistance." The term includes, therefore, not only fully-automatic translation systems that deal with text, but speech-to-speech systems and programs that involve interaction by the user, to a greater or lesser extent, including Translation Memory (TM) tools. For the purpose of this research, however, the term Machine Translation is used to refer specifically to fully-automatic MT systems that translate written data, and that require the user to do nothing more than enter some text or select a file for translation, specify the appropriate subject domain (where this feature is available) and click on the "Translate" button.

Fully-automatic MT systems for text-to-text translation can be divided into six groups, as described below. The first five categories (commercial systems) are summarised from the online Compendium of Translation Software[1] (Hutchins et al., updated July 2005).

1. MT systems for home use by the general public (some of which are little more than dictionaries);
2. MT systems for translating electronic texts on the Internet (email, web pages etc.);
3. MT systems for professional use, designed to assist professional translators (eg. off-the-shelf desktop packages, with adaptable user dictionaries and specialist topic-based glossaries, or tailor-made systems designed for particular companies or organizations);
4. Client/server MT systems designed for company intranets;
5. MT systems for company websites, providing online translation of web pages;
6. Free online systems for the translation of short pieces of text or web pages.

The types of machine translation systems selected for evaluation for this research are described in Chapter 4.

The different types of system and their uses listed above give an indication as to why we need machine translation. Six reasons why we should be interested in using computers for translation are given in (Hutchins, 2005). These are summarised below:

1. Human translators cannot cope with the huge volume of material that needs translating.
2. Repetitive, technical material is too boring for human translators.
3. Companies require consistent terminology in their translations, which computers can provide. Humans have a natural preference for variety of language.
4. Machines work more quickly than humans.
5. High quality human translation is not always required.
6. Companies with large-scale translation needs can reduce their costs.

[1] http://www.eamt.org/compendium.html

The above points clearly illustrate why research and development in machine translation is thriving, why the use of MT in appropriate contexts is increasing and, therefore, why system evaluation is so crucial.

2.2 A brief history of machine translation and its use

It is not within the scope of this thesis to provide a detailed history of machine translation. The purpose of this section is to investigate the widespread development and use of MT systems and, in so doing, to highlight the importance of MT evaluation. An understanding of how MT systems are used also serves as a starting point for the development of appropriate evaluation methods and the selection of texts to be used for this research.

Very basic machine translation systems were first developed during the Second World War, using substitution techniques based on decoding methods. However, the first public demonstration of an MT system took place in 1954, the result of a collaborative project between Georgetown University in Washington DC and IBM. This Russian-English system involved only 250 words and 6 grammar rules. In fact, the limited input made the system's capabilities seem far better than they really were, making expectations of MT unrealistically high.

More than a decade of MT research was pursued in the USA, Western Europe, the Soviet Union and Japan, until a study was funded by US government sponsors of MT. Its purpose was to advise the Department of Defence, the CIA and the National Science Foundation on research and development in the mechanical translation of foreign languages. Its findings, published in a report by ALPAC, the Automatic Language Processing Advisory Committee (Pierce, 1966), were to have a disastrous effect on MT development in the United States. The report advised against investment in further research, claiming that MT was slower, less accurate and more expensive than human translation. Instead, it recommended expenditure in other areas of computational linguistics and research into various kinds of machine-aided translation tools, such as electronic dictionaries. Three of its nine recommendations did, however, highlight the importance of evaluation. These are detailed in Section 3.3.

Following the publication of the ALPAC report, MT research in the USA came to a virtual standstill, and a negative impact was also felt elsewhere. However, in 1970, an

MT system called Systran was developed for the US Air Force, who needed to translate documents from Russian into English. This system is still in use today. In 1976, the Commission of the European Communities purchased a version of Systran to translate from English into French, and further language pairs of the European Union have subsequently been added. Today, anyone working for the European Commission, or employees of public administrations in the EU Member States, can submit texts online to be machine translated by ECMT (formerly known as EC Systran). The system translates up to 2,000 pages per hour (Petrits et al., 2001) and in 2004, it processed 693,306 pages (Angelo Torquati, personal correspondence). Although the system is used by the EC Translation Service to produce first drafts for post-editing, the main users are administrators who use it for browsing texts in a language they do not know, deciding if a document is useful enough to request a human translation, or drafting documents in another language. Today, Systran's many other clients include NATO, Ford, General Motors, Berlitz and Xerox.

Another MT success story began in 1978 when the Canadian Météo system, designed to translate weather bulletins from English into French, came into operation. In 1988 the system was extended to translate from French into English, and by 1993 it was translating around 45,000 words per day, amounting to over 16 million words per year, the equivalent of 30 person-years of translation work (Vasconcellos, 1993). Still in use today, the system produces output that requires very little human intervention, and exemplifies the success of MT for translating a sublanguage (the limited vocabulary and syntax of a particular domain).

During the 1980s many new MT projects began and novel approaches to system design were explored, particularly in Japan, where knowledge-based and interlingua-based methods were investigated. These and other approaches are described in 2.3. After years of research and development in academic institutions and government departments, MT was now attracting commercial interest. The first systems appeared on the market, particularly in Japan and the USA, and in 1981, the first translation software for personal computers became available (Hutchins, 2005).

Several tailor-made systems were designed for large corporations, who saw the time and cost-saving benefits of using MT output as a basis for polished translations. The SMART Corporation (New York), for example, began to develop tailor-made systems

for particular terminology and document types in the early 1980s; their clients include Ford and the Canadian Department of Employment and Immigration. The Pan American Health Organisation developed its own general-purpose systems in-house, to translate between English and Spanish; SPANAM has been in use since 1980 and ENGSPAN since 1985. Most of the output is post-edited to publishable quality by professional translators. LOGOS, a multilingual general-purpose system, was first used in Germany in 1983 and is still used today in many companies worldwide including the Canadian Government and AT&T in the USA. The most sophisticated system at this time, however, was METAL. This was initially a German-English system, designed to translate documents on data processing and telecommunications. The project was funded by Siemens, Munich after initial research at Texas University. Although no longer available, METAL was used by many European companies, such as SAP, Philips and the Union Bank of Switzerland.

The 1990s saw a rapid increase in the use of commercial MT and computer-assisted translation (CAT), and with the explosion of the internet, 1997 saw the launch of the first free online system: AltaVista's Babel Fish[2], powered by Systran. The goal was to eliminate the language barrier on the web (AltaVista, 1997), enabling users to translate web pages or raw text within seconds. Initially able to translate ten European language pairs, this has now increased to 36 language pairs, involving 13 European and East Asian languages. Use of the service steadily increased from 500,000 translations per day in 1998 to 1.3 million per day in 2000 (Yang and Lange, 2003), illustrating the growing need for quick translations. Babel Fish is often used as an assimilation tool, enabling users to understand the gist of a document in an unfamiliar language; it is used to disseminate information (some web sites provide a link to Babel Fish in an attempt to reach a wider audience); it is used as a communication tool between speakers of different languages and also as an aid for language learning. Following the appearance of Babel Fish, many other online MT engines have since emerged, and the number of available language pairs is growing all the time to reflect demand.

During the 1990s, MT research focused increasingly on corpus-based methods, such as example-based machine translation (EBMT) and statistical machine translation (SMT), beginning with IBM's Candide project (Brown et al., 1990). Both approaches are described in the following section. These methods have not yet had a great deal of

[2] http://world.altavista.com/

commercial impact, as research continues and development is still at a relatively early stage. However, one company to eventually commercialise a statistical approach is Language Weaver, founded in 2002. The company claims that its statistical MT modules can translate up to half a million words per minute from multiple users on the same network (Language Weaver Press Release, 2005). Working on servers or desktops, systems can be tailor-made for governmental and commercial use.

A second example of a corpus-based approach is MSR-MT, Microsoft Research's data-driven system, trained on over a million bilingual sentence pairs per target language, extracted from translation memories and glossaries. The system's primary function was to translate Microsoft's product support services knowledge base from English into French, German, Japanese and Spanish. Plans now involve extending the system to translate into Italian, Chinese, Korean and other languages, and it is envisaged that tens of millions of dollars in translation services will be saved each year by using this new system (Richardson, 2004).

During the last decade, the globalisation of business, along with the use of the Internet to advertise and sell products worldwide, has created an enormous need for quick automated translation. Human translations are costly and time-consuming, making MT an invaluable support tool or a viable alternative for the translation of certain document types for particular purposes. According to (Hutchins, 2002) "The aim of using computers for translation is not to emulate or rival human translation but to produce rough translations which can serve as drafts for published translations, as gists for information gathering, and as cross-language communication aids."

A number of translation memory (TM) systems are now compatible with MT (eg. Trados Translator's Workbench can be used in conjunction with Systran and LOGOS) and many companies have started to incorporate MT in the translation workflow. For instance, Bowne Global Solutions, recently taken over by Lionbridge Technologies, Inc., has adopted a common methodology known as Maestro, for integrating MT, human translation and TM for large repetitive translation projects. First, a Trados translation memory database (containing previously translated sentences aligned with their human translations) is searched and all matching segments over a given threshold value are inserted into the target text. Segments not found in the TM are then automatically machine translated and marked as such. This draft translation is post-

edited and the translation memory is finally updated in preparation for the next job (Walker, 2004).

Today, new versions of systems like Systran, which were originally developed for mainframe computers, can now be bought off-the-shelf as desktop applications. Systems have become less expensive and now that MT is more affordable, smaller companies and freelance translators have the opportunity to make savings in terms of both time and money. In one controlled study, for example, post-editing machine translations of three marketing brochures (2,470 words in total) took slightly less than half the time required to translate the same texts from scratch. Where translation took 6 hours 44 minutes, post-editing of MT output took only 3 hours 11 minutes (Guerra, 2003).

Detailed histories of MT research, development and use can be found in (Hutchins 1986, 1992 and 2000). In addition, the regularly updated Compendium of Translation Software (Hutchins et al., 2005) available from the European Association for Machine Translation (EAMT) website[3] lists over 80 pages of commercial products, including MT systems currently handling 44 languages.

2.2.1 Concluding remarks

It has been observed that the rapidly growing demand for quick and cheap translations has led to the increased development and use of MT throughout the world. A large number of companies and governmental institutions are using tailor-made and off-the-shelf, PC-based systems for particular purposes, text types and domains, and the use of free online MT is still growing. Furthermore, we are now seeing the globalisation of MT research, with the development of systems to translate between more language pairs to meet user needs.

Gaining an understanding of who uses MT and for what purpose has not only provided a context for the current research, but has also formed a basis for the development of evaluation strategies to suit real MT scenarios for this work.

2.3 How do machine translation systems work?

This section takes a brief look at how fully-automatic machine translation systems work, introducing some of the reasons why translation is difficult for computers and

[3] http://www.eamt.org/

why errors in MT output occur. System architectures can be divided into several groups, as shown below, but as (Hutchins, 1995) points out, the differences between system types are becoming less useful for categorisation. For example, transfer systems can incorporate features of interlingua systems (and vice-versa), rule-based systems are using probabilistic data and corpus-based approaches are integrating rule-based methods. Furthermore, some research has focused on knowledge-based approaches (KBMT) involving attempts to encode real world knowledge to make systems more intelligent by making more appropriate lexical choices in particular contexts.

2.3.1 Rule-based approaches

Rule-based approaches were the main focus of MT system development until the end of the 1980s, and these methods are still being used today. The **direct approaches** of the first generation of MT systems were developed for one particular language pair. Source language sentences are converted directly into target language sentences using nothing more than a large bilingual dictionary and a program for analysing the source language and generating the target language. The analysis phase involves some identification of morphology, in that words are reduced to their uninflected forms before dictionary look-up. The text is then translated more or less word-for-word, and limited rules are applied to rearrange words to suit the syntax of the target language before the translation is generated. For instance, a simple parse enables the system to reposition adjectives after nouns as is often required when translating from English into French. These limited rules are quite insufficient, as they do not cater for the vast number of structural variations between source and target languages.

The inadequacies of the direct approach led to the development of the second generation, indirect **linguistic knowledge systems.** Indirect methods, involving a kind of intermediate representation of meaning between source text analysis and target text generation have two main approaches. The **interlingua approach** involves two stages; first the source text is analysed to produce an abstract, language-independent representation of its meaning and second, this information is used to generate a target text without, in theory, being influenced by the original. The intermediate representation is known as an 'interlingua', being neutral between the source and target languages. There are, however, problems with this approach, in that designing a true abstract representation of language remains an extremely difficult goal for linguists, so it is very hard to generate a target text with no influence from the source. The main advantage of

this method over the direct approach is that the interlingua can be used for any language pair, so adding new languages to a system involves only the addition of analysis and generation modules for the languages concerned.

The second indirect method is the **transfer approach**, which involves three stages; analysis, transfer and synthesis (or generation). First the source text is analysed and converted into a source language-oriented representation. Second, a bilingual transfer module converts this into a target language-oriented representation and finally, this is used to generate the target text. This type of system contains a number of components; monolingual dictionaries for the source and target languages, containing morphological, grammatical and semantic information; a bilingual dictionary and grammars and components for lexical and structural transfer. One problem with this approach is that when a new language is added, a transfer module is required for every language pair involving that language, in addition to new analysis and generation modules. In spite of this, transfer methods are preferred to interlingua approaches, due to the difficulties of creating language-independent representations, which are required by the latter.

2.3.2 Corpus-based or data-driven approaches

This third generation of machine translation systems, involving example-based and statistical-based techniques, involves the use of large corpora of sentences aligned with their human translations. The hardware required for data-driven approaches has only recently become affordable, but some systems are now in use. Two examples are given in Section 2.2.

The idea of **Example-Based Machine Translation** (EBMT) was first proposed by Makoto Nagao in 1981, but was not published until three years later (Nagao, 1984) and serious research did not start until 1989. EBMT is essentially translation by analogy: a translation is generated at run-time by comparing word sequences, rather than single words, in the source text against a large parallel aligned corpus. The closest matching fragments are selected and their translations recombined to form the target text. Additionally, EBMT systems may incorporate techniques from rule-based and/or statistical methods. For example, systems may use bilingual dictionaries or monolingual thesauri.

A major advantage of this approach is that linguistic knowledge (in terms of word order, agreement etc.) can be automatically captured from examples in the corpus. Furthermore, it is argued that EBMT systems are easily improvable, by adding more sentence pairs to the corpus, in contrast to rule-based systems, which require the analysis, modification and addition of complex rules. One problem with EBMT, however, is that there can be several different translations of the same source text fragment, some of which may be more appropriate than others. Furthermore, it can be difficult for the program to automatically select which fragments of a parallel segment actually correspond to one another.

Research into **Statistical Machine Translation** (SMT) began with work by (Brown et al., 1990) on the Candide system at IBM. As (Somers, 1999) points out, a major distinction between EBMT and SMT is that where example-based approaches use a bilingual corpus as a main knowledge base at run-time, statistical machine translation methods rely on corpus-driven probabilities, which are computed in advance. Like EBMT, SMT relies on a bilingual aligned parallel corpus, but it is distinctly different. The corpus is first aligned at the sentence-level, then at the word-level. Probabilities are then computed that any given word in a source text segment corresponds to zero, one or more words in the aligned translation. The result is a list of translation possibilities for each source language word with the computed probability of each target language word (or group of words) being a correct translation. This is known as the 'translation model'.

During the translation process, the computed probabilities are combined and the highest scoring combination is selected as the translation. A monolingual target language model containing probabilistic information on word co-occurrence is then used to reorder the target language words to generate sentences. The most successful systems now use 'phrase-based methods', whereby sequences of words are translated together (Och, 2002; Zens et al., 2002; Koehn et al., 2003; Vogel et al., 2003; Tillman, 2003).

A major problem with the statistical approach is that a huge amount of data is required to gather reliable statistics from which probability information can be obtained. Furthermore, the high-quality data required can be expensive. According to (Bennett and Gerber, 2003), one million bilingual sentence pairs is a good size for a training set for a general purpose MT system. Plenty of memory and powerful processors are therefore required to translate in real time.

2.3.3 Concluding remarks

This overview of MT system architectures has provided an insight into some of the complexities involved in the automatic translation of natural language. Irrespective of the system architecture, translation is difficult for computers, and MT systems continue to produce imperfect output of varying quality. The following section addresses particular problems encountered by MT engines and helps to explain why particular errors in output occur.

2.4 Why is translation difficult for computers?

Fifty years ago, people assumed that automatic translation would eventually replace human translators. However, during the development of the first rule-based MT systems, linguists began to fully understand the complexities of describing and encoding the components of entire languages. Not only are enormous dictionaries required for these systems (bearing in mind that a person's working vocabulary in English comprises up to 100,000 lemmas and the Oxford English Dictionary contains around half a million root words), but also complex grammatical rules. Translating from one language into another is not merely a matter of word-for-word substitution. (Arnold, 2003) makes the following observation: "Part of the reason why translation is difficult for computers is that translation is just difficult: difficult even for humans." He reminds us that translation is a creative task and that computers simply follow rules, mechanically and literally.

At this point, it should be noted that since few commercial corpus-based systems are currently available, the focus of this research, and of Arnold's discussion, is on rule-based systems. Arnold divides machine translation problems into four categories based on the point at which they occur during processing by rule-based systems. His four headings are duplicated below and enhanced by examples of errors observed during the analysis of MT output from the four systems used for this research. (Details of these rule-based systems are provided in Chapter 4.) For the sake of completeness, the nature and causes of errors in output from corpus-based MT systems are discussed in 2.4.5.

2.4.1 The analysis problem

This problem occurs when the source text contains words or structures, which are either not understood by the analysis component, or can be parsed in more than one way. Where a source language word or phrase has two or more different meaning

representations, the system must arrive at one decision. For a machine translation system, two kinds of ambiguity exist; lexical ambiguity occurs when a source language word has more than one translation, and structural ambiguity arises when the system identifies more than one way to parse a sentence (eg. to relate an adjective to one or more nouns, an adverb to an event, a pronoun to its referent, etc.).

By using contextual information, a human can normally parse a sentence correctly, but MT systems only work at sentence level. According to (Pinker, 1994), "Computer parsers are too meticulous for their own good. They find ambiguities that are quite legitimate, as far as English grammar is concerned, but that would never occur to a sane person." Pinker gives the following example from an experiment with one of the first computer parsers in the 1960s: the computer succeeded in parsing the same sentence in five different ways.

Time flies like an arrow.

(1a) Time proceeds as quickly as an arrow proceeds. (Intended meaning.)

(1b) Measure the speed of flies in the same way that you measure the speed of an arrow.

(1c) Measure the speed of flies in the same way that an arrow measures the speed of flies.

(1d) Measure the speed of flies that resemble an arrow.

(1e) Flies of a particular kind, time flies, are fond of an arrow.

This is, of course, an extreme example, but two alternative parses for some part of a sentence are not uncommon. The examples in (2a) and (2b) from software user manuals show erroneous French-English machine translations resulting from incorrect parsing.

(2a)	French source:	**Coordonnées** du site d'observation
	English MT :	**Coordinated** of the observation site
	Literal translation:	**Coordinates** of the site of observation
	Human translation:	Observation site **coordinates**
(2b)	French source:	La liste commence par une **case** à cocher…
	English MT:	The list starts with a **puts** to check…
	Literal translation:	The list starts with a **box** to check…
	Human translation:	The list starts with a check **box**…

In (2a) the meaning is obscured in the MT output because 'Coordonnées' has been labelled as an adjective (or past participle of a verb) rather than a noun. Both parts-of-speech can be applied to this word in French, but the engine has made the wrong choice in this instance. As the phrase is a section heading, the article has been omitted in the French source text, which would otherwise have led the system to make the correct choice. In (2b), the word 'case' has been processed as a verb instead of a noun, rendering this part of the machine translated phrase incomprehensible.

For a human translator, real world knowledge normally enables an immediate, correct interpretation of a source text with no thought of ambiguity, but for an MT system a huge amount of work is required to encode even a small amount of this knowledge.

2.4.2 The transfer problem

This problem is caused by the fact that languages can use very different structures to express the same content, so rules are required to relate the source and target representations. For example, in (3), the MT system has translated the multiword verb construction literally. We assume that no rules have been encoded for translating this French structure into English.

(3)	French source:	Les interdictions **sont plus prioritaires que** les autorisations.
	English MT:	Bans **are more priority than** authorisations.
	Human translation:	Restrictions **take priority over** authorisations.

In the construction in (4) below, the French verb 'permet' does not require an indirect object. However, the English translation requires this for the sentence to read fluently. A human translator would decide on the pronoun 'you' without difficulty, and would prefer to use the verb in the infinitive, but providing sufficient knowledge to an MT system for it to make this choice would be extremely complex.

(4)	French source:	Cette fonction **permet de quitter** l'atlas.
	English MT:	This function **allows leaving** the atlas.
	Human translation:	This function **allows you to exit** the atlas.

As in (4), problems often occur when information in one language is not specified. This can result in either unnecessary or omitted words in the machine translation output. In Japanese, for example, information on whether a noun phrase (NP) is singular or plural, definite or indefinite does not have to be provided. In such cases, the MT system has to arrive at a decision on which information to insert to meet the requirements of the target language. In Italian and Spanish, subject pronouns are often omitted, so without such an indicator, an Italian verb form such as 'è', and equally the Spanish 'es' can translate into 'he is', 'she is', 'it is' or 'you are'. Where a human is able to use contextual information from anywhere in a text to achieve the correct interpretation, an MT system can only search for information within the same sentence, such as adjectival agreement (which is not always available), in order to produce a correct translation. In many cases, these kinds of ambiguities can only be resolved by using information from adjacent sentences, something which is beyond the ability of current MT systems.

2.4.3. The synthesis problem

The synthesis problem occurs when the same content can be expressed in several ways. Sometimes a source language word has two or more legitimate translations with very different meanings, such as in (5a). In some instances, on the other hand, one word choice may be preferable to another in a given context. For example, some companies prefer to use particular terminology, and in such cases, company-specific dictionaries can be developed to ensure that the preferred translations of these words are consistently generated. In other cases, a correct semantic translation may be inelegant due to a co-occurrence of similar sounds or words (see 5b) or a collocation that appears unnatural to a human (5c, 5d). The examples below are again machine translated from French.

	MT	Suggested alternative
(5a)	dry ink in the **buzzards**	**pipes**
(5b)	See the websites **listed** on the list	shown, provided etc.
(5c)	cut and **stick** the text	**paste**
(5d)	from the **principal** menu	**main**

For (5b) a better option may be to omit the verb all together, but rule-based systems do not have these decision-making capabilities.

2.4.4. The problem of description

This problem concerns the collection and description of the knowledge required by an MT system. The system must have sufficient knowledge of the lexical, morphological, syntactic and semantic rules of a given language. Regular patterns are easy for computers to understand, but the many exceptions to the rules and the way in which different rules interact with one another makes language description for computer processing incredibly complex.

As mentioned at the beginning of 2.4, huge dictionaries are required for each language, and any word not found in the dictionary is left untranslated in the target text. Although many commercial MT systems allow new entries to be added to a user dictionary to supplement or override translations in the system dictionary, entering all the specialised vocabulary required by a particular organisation is an enormous task. Noun strings in particular can produce unnatural translations when the whole multi-word unit is not found in the dictionary. For example, where in English we tend to group strings of nouns together, in French, prepositions are also required to link the nouns. Examples of French noun strings with disfluent English machine translations are shown in (6a) and (6b).

(6a)	French source:	La barre des menus
	English MT:	**The bar of the menus**
	Human translation:	The menu bar

(6b)	French source :	Propriétés de la connexion internet
	English MT:	**Properties of the internet connection**
	Human translation:	Internet connection properties

2.4.5 Errors generated by data-driven MT systems

One main advantage of data-driven approaches is that some of the problems concerning disambiguation, pronoun resolution and the mistranslation of idioms, which typically occur in output from rule-based systems, are less likely to occur, as some linguistic knowledge is automatically captured from corpora. However, there is still a high level of similarity between errors identified in output from rule-based and data-driven systems, although the causes are quite different.

Lexical errors are found in output from corpus-based MT systems, largely due to source language words, which have multiple legitimate translations in the aligned data. Furthermore, the larger the corpus, the more likely these errors are to occur. Lexical errors may also appear when the system has to rely on a dictionary in cases when source language words are not found in the corpus. Attempts can be made to reduce these problems, by prioritising translation equivalents using additional rules or by using a corpus in a domain, which is specific to the translation task.

Grammatical errors also occur in output from data-driven MT systems. In EBMT, for example, the target language model works with a fixed-size window, even though more distant words can be grammatically related. For this reason, incorrect verb inflections can occur when they are distant from their subjects. This type of error is far less likely to appear in output from rule-based MT systems.

Syntactic errors, unnecessary and omitted words can also be found; these are often due to corpus alignment errors. Similarly, disfluencies can occur when partial translations are pieced together, even in cases when the individual parts are correct. This is often referred to as a "boundary friction" problem. A German-English example is shown in (7) below.

(7) Der Junge aß **den Apfel.** The boy ate **the apple.**
(Use of accusative case **(den)**, required for the direct object **(Apfel)**
But:
Den Apfel war rot. **The apple** was red.
(Incorrect use of translation fragment: German nominative case **(Der)** is required here as **Apfel** is now the subject of the sentence)

2.4.6 Concluding remarks

We have seen that for a human to translate accurately from one natural language into another, he/she must not only have the ability to understand the source language, but also to convert the meaning into the target language while adhering to TL rules and conventions. Having explored the complexities of natural language understanding and generation, we can better appreciate the reasons why translation is so difficult for

computers. The different possible interpretations of source text phrases, the structural differences between languages, the existence of multiple legitimate translations of some words and the sheer size of the lexicon are some of the main causes of errors in MT output. Furthermore, a lack of real world knowledge and an inability to benefit from contextual information beyond the sentence level do not help to solve these problems.

This investigation of language processing difficulties has enabled us not only to understand the causes of errors made by MT systems, but also to predict the kinds of errors that are likely to occur. These include untranslated words, incorrect translations of words (in terms of meaning or contextual appropriateness), omitted and unnecessary words and a number of different structural errors. A detailed classification of error types, developed specifically for this research, is presented in 4.5.

2.5 Background conclusions

This chapter began by providing a number of reasons why MT is needed and an overview of the growing use of MT systems for a variety of purposes and language pairs. This overview has served three purposes; first, it has provided a wider context for the current research; second, it has enabled us to understand how widely used MT systems are throughout the world and, therefore, why the evaluation of MT output, for the various stakeholders involved, is so important; third, an understanding of real MT scenarios serves as a starting point for the development of appropriate evaluation strategies for this research.

The second part of this chapter focused on the design of MT engines, the translation difficulties encountered by rule-based systems and the kinds of errors we can expect to find. As mentioned in 2.4.4, MT output can be improved to a certain extent by adding terms and their preferred translations to user dictionaries, where this function is available. However, the main purpose of this feature is to improve particular lexical choices and there is only a very limited ability to reduce structural errors (eg. noun string word order). The fact remains that, even with the aid of user dictionaries, particular errors in MT output persist.

Having considered some of the kinds of errors likely to occur in MT output, it is evident that some form of classification scheme can be developed, whereby errors can be categorised according to type and how they might be automatically identified. Using

such a classification scheme to develop a suite of error detection algorithms may provide a reliable evaluation method for both developers and end-users, provided that scores correlate highly with human judgements,

Before developing such a method, however, an investigation of existing MT evaluation methods, both human and automated, is required. First, we need to select appropriate methods for the human evaluation of MT output in order to obtain judgements against which scores from experimental automated methods can be compared. Second, we need to investigate existing automated methods to gain an understanding of progress already made in the field and to provide directions for this research. These issues will be examined in Chapter 3.

Chapter 3

MT evaluation methodology: literature review

During the last five decades, a vast amount of research has been published, reports written and projects organised, all contributing to progress in MT evaluation methodology. This chapter has two main objectives. The first is to explore and present a critical analysis of the different methods used for the human evaluation of MT output; an understanding of the various different approaches will enable the selection of the most appropriate method(s) and resources to use for this research, to obtain human judgements of MT quality against which scores from a novel automated approach will later be compared. Furthermore, such an analysis will enable us to gain an understanding of why the human evaluation of MT output is so complex, costly and time-consuming and, therefore, why interest has turned more recently to automated methods. The second objective is to explore and critically analyse existing approaches to automated evaluation to gain an understanding of their strengths and weaknesses and to enable us to identify areas for improvement and further research.

There are, of course, many ways to evaluate MT systems besides judging the quality of their output; by way of an introduction to MT evaluation, this chapter begins in 3.1 with a brief overview of the different ways in which MT systems need to be evaluated. This is followed in 3.2 by a look at some of the difficulties involved in the evaluation of MT output. A number of published recommendations on MT evaluation methodology, resulting from various projects and reports, are presented in 3.3, and in 3.4 different methods for the human evaluation of MT output are categorised, described and critically analysed. Automated evaluation methods are similarly explored in 3.5.

3.1 Types and purposes of MT evaluation

The design and development of machine translation engines is complex and, as a result, systems can be expensive to buy. With large investments at stake, evaluations before, during and after development are required for different purposes by the many groups of people involved. Particular kinds of evaluation are essential for investors, researchers, developers, vendors, project managers and end-users, such as translators, post-editors

and monolingual administrators who need to translate documents quickly for information purposes. Types of MT evaluation are shown in Table 3.1.

Type of evaluation	Description	Purpose	Stakeholders
Feasibility evaluation	Evaluation of the potential of a new MT approach, eg. by testing a new approach to a translation problem specific to a particular language pair	To measure the feasibility of investment into further research and development of a particular approach	Investors Researchers Developers
Requirements elicitation	Prototypes are developed to demonstrate specific functions for possible implementation as part of an MT system	To elicit reactions from potential stakeholders to guide further development	Investors Developers Project managers End-users
Internal or Progress evaluation	Regular evaluations of MT components prior to system release, eg. before and after modifications. A "glass box" approach: evaluations are performed with access to the internal workings of the system. Test suites are often used to evaluate the successes or failures of new translation rules	To test the linguistic coverage of the system. Eg. developers need to know that a new grammar rule will work successfully in all circumstances To check that modifications which should improve the system, do not have an adverse effect elsewhere in the software (iterative testing)	Investors Researchers Developers
Diagnostic evaluation	Evaluation of functionality characteristics of prototype by researchers/developers. Glass-box approach	To diagnose reasons for unexpected results produced by a system	Developers
Declarative evaluation	Evaluators judge MT output quality using selected metrics. The quality of the output may be evaluated at different stages of a system's maturity (eg. before and after dictionary update), or translations from different systems can be compared (see Comparison evaluation). Black box approach: access to output only	To measure how well a system translates for the needs of a particular end-user, eg. in terms of fidelity, fluency etc.. Findings indicate whether buying a system would be cost-effective by saving time for translators or post-editors	Investors Researchers Developers Vendors Project Managers End-users
Usability evaluation	Evaluators representative of end-users test how easy the application is to use. Researchers may devise questionnaires on system usability, or record how long it takes for subjects to complete particular tasks using the software	To measure how useful the product will be for the end user To evaluate the user-friendliness of the interface	Investors Researchers Developers Vendors Project Managers End-users
Operational evaluation	Managers calculate the purchase and running costs of an MT system and compare these with its benefits	To determine the cost-effectiveness of a system in a particular operational environment	Investors Researchers Developers Vendors Project managers
Comparison evaluation	Declarative, Usability and Operational evaluations may be used to compare systems. When performing a comparison evaluation, the same attributes must be tested for each system, using the same criteria and metrics for reliable results	To help users decide which system will best suit their needs.	Investors Researchers Developers Vendors Project Managers End-users

Table 3.1 Types of MT evaluation

A good deal of research has gone into defining and categorising the different kinds of MT system evaluation required. Table 3.1, based on the MT evaluation glossary on the FEMTI website[1] (2005) and publications by White (2000, 2003), EAGLES (1996), Arnold et al. (1993), and Van Slype (1979) summarises the types of MT evaluation required, and for whom they are of interest.

The current research is concerned with two of the categories shown in Table 3.1: declarative evaluation and comparative evaluation. Our focus is on the evaluation of MT output from competing systems, using a "black box" approach: evaluations are based on judgements of the quality of the output alone, without any access to the internal components of a system or any interest in the causes of errors; issues which are of particular concern for developers. Nevertheless, the black box evaluation of MT output is still useful for all stakeholders, including developers, who need to monitor improvements or deterioration in output quality as translation rules are modified. Furthermore, black box evaluation is entirely platform-independent, enabling the comparison of different MT systems, regardless of their architectures. The focus now turns to declarative evaluation, and first, to some of the complexities involved.

3.2 The difficulties of evaluating MT output

The evaluation of any translation, whether produced by a human or a machine, is complex and, as a result, it can be expensive and time-consuming. The difficulties of MT evaluation described below help us to understand why, in the last five years or so, there has been a growing interest in quicker and cheaper automated evaluation methods, which attempt to compute scores that reflect human quality judgements.

Translation quality is difficult to quantify and even when some form of description and measurement is developed for á given quality characteristic (eg. fluency or fidelity), it can be open to different interpretations. Furthermore, translation is legitimately subject to lexical, stylistic and other variation. MT evaluation relies on either the objective scoring of specific linguistic phenomena using test suites (a method mostly used by developers), or the subjective quality judgements of evaluators, who score individual fragments of text using particular metrics, with or without the source text or information on elements of meaning that should be contained in the translation.

---

[1] http://www.isi.edu/natural-language/mteval/

Some of the difficulties involved in the evaluation of MT output are described below. Examples of ways in which these problems have been reduced or overcome are described in the review of human evaluation methods in 3.4.

3.2.1 Translation quality: interpretation and subjectivity

A major problem for MT evaluation is that translation quality is difficult to define and quantify and, like judging paintings in a competition, comparing the quality of language in different texts can be equally subjective. According to (House, 1977), evaluating the quality of a (human) translation means judging it in terms of two sets of standards: those based on the source text and culture and those based on the target language and culture. The first concerns faithfulness to the original text (content, style, function or intention and form) and the second concerns the degree to which the translation conforms to the norms of the TL and culture. This includes grammaticality, acceptability and the situational appropriateness of the translation. These quality components are also applicable to machine translations. However, interpreting, quantifying and measuring these quality attributes is a complex task.

For (McClure and Flanagan, 2003) "the difficulty of assessing and improving translation quality is MT's most intractable problem. Translation quality is inherently subjective and therefore difficult to measure." Just as translators have their own styles and preferences, readers have their own opinions on the quality of a translation. Evaluator judgements are based on a range of factors, including their interpretation of the quality measurement, their own exposure to, use and experience of the target language (both written and spoken) and their knowledge of the relevant subject matter. This means that even in a simple exercise, such as arranging translated sentences from "best" to "worst", the judgements of one evaluator can differ from those of another. The same evaluator may even have a different opinion on a different day, which can be due to any number of reasons, including boredom, tiredness, hunger, personal and other circumstances.

3.2.2 The absence of a "gold standard"

The evaluation of a product is normally designed to measure its attributes against a "gold standard". However, evaluating the quality of a translation, whether by a human or a machine, is not so straightforward because no perfect standard exists. A candidate

for comparison against MT output would be an expert human translation, but if a number of experienced translators were asked to translate the same text, it is unlikely that any two versions would be the same. Each translator would have his/her own interpretation of the source text, would gravitate to particular word choices and display an individual style, albeit within the parameters of the genre. Provided that all translations were free of grammatical errors, read like natural English, in a style appropriate to the target text genre, and conveyed the same information as the source document, deciding on the "best" translation would be difficult or impossible.

Of course, not all evaluation methods require a human translation for comparison. Evaluations of fluency, for example, require access only to the MT output. However, methods which do rely on the use of some human translation (eg. evaluations of fidelity) must take into account the fact that no single translation is the only correct version.

3.2.3 The need for bilingual knowledge or human translations

In order to rate the fidelity of a machine translation, an evaluation requires the expense of (a) recruiting bilingual evaluators or (b) acquiring human translations.

In the first case, bilingual evaluators (native speakers of the target language with a good knowledge of the source language and subject matter) can be used to compare the content of the MT output against the source text. Evaluators are normally required to assign a score to each fragment of text according to how much of the original content they perceive to be conveyed in the translation. Bilingual judges can, of course, be hard to find and are likely to be more expensive to employ than monolingual subjects.

In the second case, human "reference translations" (expert translations, conveying all the factual information of the original without any stylistic flourishes) are produced, against which evaluators with no knowledge of the source language compare and score the content of the MT output. Producing reference translations is very costly and time-consuming, particularly as declarative evaluations normally require the analysis of a large number of sentences, to give a clear picture of system performance.

3.2.4 The training effect

The 'training effect' can occur when evaluators see two or more different translations (eg. from different MT systems) of the same source sentence or text. This can call into question the validity of scores for particular parts of a translation, as an evaluator may perceive a sentence to be more intelligible or more faithful to the original than it really is (and score it more highly) if he/she has seen a different translation of the same sentence before. Similarly, a string of particularly bad sentences might make a reasonable subsequent sentence appear better than it is.

3.2.5 Different uses for MT output

Since expectations of MT systems have become more realistic, and perfect quality is no longer expected, MT users have found a number of tasks for which imperfect MT can be useful. Raw output is sufficient for a variety of text-handling tasks, and can be particularly useful for people with no knowledge of the source language. A number of tasks that might be carried out using raw MT output is provided in (Taylor and White, 1998). Their list includes filtering (discarding irrelevant documents), detection (finding documents of interest, perhaps with a view to requesting a human translation), triage (ranking documents in order of importance), information extraction, gisting (producing a summary of a document) and publishing (by post-editing the text to an acceptable quality).

The variety of different uses for raw MT output means that certain evaluation methods are more appropriate in particular scenarios. For instance, a company planning to use MT for nothing more than filtering is likely to be more interested in fidelity than fluency when comparing systems. The level of granularity required in the scoring of MT output will also vary depending on user requirements. Furthermore, some purchasers may prefer a lower quality, less expensive system if an evaluation concludes that its output is of sufficient quality to perform required tasks. In short, a single evaluation approach is never appropriate for all scenarios.

3.2.6 Tackling the difficulties

On the whole, generally accepted methodologies have emerged to tackle the above problems, even though no common standard for evaluation exists or is, in fact, appropriate for all scenarios. For instance, the problem of subjectivity is addressed by obtaining judgements of the same texts from several evaluators to calculate a mean

score, and the reliability of results can be increased by using a large number of texts. Legitimate translation variation and, hence, the lack of a 'gold standard' means that some approaches, such as the automated BLEU method (Papineni et al., 2001), require multiple human 'reference translations' against which MT is automatically compared. The training effect can also be avoided by collating texts in such a way that no evaluator sees more than one translation of the same source sentence. These solutions do of course incur additional expense.

Examples of ways in which the above problems have been addressed in various different human evaluations are described in Section 3.4 and, in addition to these solutions, a number of recommendations for MT evaluation have been made in various published reports. Some of these recommendations are informative for the current research and are highlighted in the next section.

3.3 Recommendations for MT evaluation

As described in 2.2, after the first decade or so of MT research, the ALPAC report (Pierce, 1966), had a detrimental effect on MT research. The report did, however, encourage work on evaluation in three of its nine recommendations:

> "Work should be supported on such matters as
>
> - practical methods for evaluation of translations;
>
> - evaluation of quality and cost of various sources of translations;
>
> - evaluation of the relative speed and cost of various sorts of machine-aided translation;"

In terms of practical methods for the evaluation of translations, the report also provided a detailed description of JB Carroll's evaluations of intelligibility and fidelity. Carroll's methods and findings still have value today, and a number of subsequent evaluations have adapted his approaches and metrics, which are described in Section 3.4.

More specific recommendations were published in 1979, when the European Commission needed guidelines for conducting ongoing evaluations of its own MT systems. The result was the publication of an extensive report: *Critical Methods for Evaluating the Quality of Machine Translation* (Van Slype, 1979). The aims of the study were to present an outline of methods for evaluating MT (whether practiced or

proposed), to provide a critical analysis of each method and, as a result, to advise the Commission on MT evaluation methodology. The report made a major contribution to the field, as it provided dozens of examples of evaluation methods and made recommendations based on thorough research. The report identified seven facets crucial to the evaluation of human and machine translations (pp. 12-14), all of which were to influence subsequent research. These are summarised below:

1. The aims of an evaluation must first be defined;
2. A definition of translation quality must be established for each evaluation;
3. A text typology is needed, so that particular text categories can be associated with particular translation methods;
4. Effective and efficient criteria are required to measure translation quality;
5. Criteria and methods for macro-evaluation (black box) must be analysed for cost-effectiveness;
6. The most effective micro-evaluation (glass box) methods for improving the system must be identified;
7. Texts and evaluators must be carefully selected to make the evaluation valid and cost-effective.

Van Slype stressed the importance of identifying who an evaluation is for and what they expect from it, before deciding on the evaluation criteria. Additionally, as translation quality is hard to define, and MT output cannot achieve the same quality as human translation, criteria must be chosen according to specific user requirements.

The issue of cost-effectiveness led Van Slype to emphasise the importance of limiting the evaluation criteria to the essential minimum. He noted that Carroll had found a strong correlation between judgements of intelligibility and fidelity, and recommended that if a strong correlation is found between scores from two independent metrics, then one method should be sufficient. This idea of finding correlations between scores for different attributes of a translation has had a great impact on MT evaluation. For example, researchers have used reliable results from earlier evaluations to validate or

reject scores obtained from new methods. More recently, this methodology has played a major role in the validation of new automated MT evaluation methods.

Having critically assessed a number of different evaluation methods, Van Slype made a number of recommendations relating to the design of declarative evaluations. These are summarised below:

- Texts for an evaluation should amount to between 5,000 and 10,000 words, comprising passages of 5 to 20 sentences from 20 to 40 documents;
- Different text types (excluding those not suited to MT, such as literary works) of varying difficulty should be used. They should be related to fields covered by the system's dictionaries;
- The number of evaluators depends on the subjectivity of the criterion, and whether they are paid or unpaid. Between 4 and 10 evaluators are suggested for intelligibility;
- Evaluation criteria depend on the user, who may not require a perfect translation;
- Criteria must be valid, reliable, applicable to both MT and human translation, sensitive and efficient (reliable at a minimum cost);
- An effective macro-evaluation should take into account intelligibility, fidelity, consistency, usability and acceptability;
- A 4-point scale is preferable when scoring sentences for their intelligibility and fidelity;
- If cost is a problem, evaluating intelligibility is the most efficient method because only the target text is required;
- Evaluating correction (post-editing) times is a good way to evaluate a system at different stages of maturity.

Van Slype's recommendations, based on the analysis of many different evaluation methods, were useful not only to the European Commission, but to many stakeholders in the field of MT throughout the world.

Further recommendations were made in a report by the Japan Electronic Industry Development Association, entitled *JEIDA Methodology and Criteria on Machine Translation Evaluation* (Nomura and Isahara, 1992). In agreement with (Van Slype, 1979), the report stressed the importance of judging systems according to their context of use. This project involved devising questionnaires to determine the needs of particular end-users, managers, researchers, developers or investors so that specific user requirements could be matched against a particular type of system.

After years of experiments with criteria, scales and metrics, the MT community identified the need for a set of standards and guidelines for MT evaluation. The EAGLES initiative (1993-1999) was set up by the European Commission to propose standards, guidelines and recommendations for good practice in the evaluation of language engineering products in general. Taking into account the needs of users and the wide range of products available, the group recognised that different types of evaluation would be required for different kinds of software: the evaluation of MT systems was investigated as part of the subsequent ISLE project (International Standards in Language Engineering), funded by the European Union, the National Science Foundation of the USA and the Swiss Government. The aim was to develop a taxonomy of quality characteristics for MT systems, and the appropriate measures to use when evaluating them. It was designed to help anyone interested in MT evaluation to select criteria according to their own needs. The taxonomy developed progressively in response to feedback from regular workshops. The resulting Framework for the Evaluation of Machine Translation in ISLE (FEMTI) website[2] offers information about the project and presents the taxonomy along with suggestions on evaluation methods (both human and automated) and appropriate references to published evaluations. Some of these evaluations are described in 3.4 and 3.5.

Having gained an understanding of the difficulties involved in translation evaluation and knowledge of published recommendations and the reasons behind them, the purpose of the remainder of this chapter is to explore and critically analyse methods used to date for declarative MT evaluation. First, an investigation of human evaluations in 3.4 will enable the informed selection of one or more appropriate methods to adopt for this research, the aim being to acquire reliable human quality judgements of translations of

[2] http://www.isi.edu/natural-language/mteval/

texts from a new corpus, against which scores from an experimental automated method can be compared. Furthermore, this study will guide decisions on, inter alia, the number of evaluators and texts to use, evaluator training, the wording of metrics and the coordination of results. Second, an investigation of automated evaluation methods in 3.5 will enable us to establish progress already made in this area. This will enable an understanding of the advantages and drawbacks of current approaches, and will provide directions for this research.

3.4 A critical analysis of human evaluation methods

Human declarative evaluations have involved a wide variety of approaches, and descriptions of various methods are presented in (Van Slype, 1979), (White, 2003), (Hovy et al., 2003) and (King et al., 2003). It would, however, be impossible to describe every type of evaluation here. In fact, some published methods provide insufficient detail, in terms of the number of texts, evaluators, metrics and results, for an evaluation to be replicated or considered as a possible approach to use for this research. Some methods were devised but never actually applied, and many examples of these can be found in (Van Slype, 1979).

The methods described here represent a number of quite different approaches that:

a. have been designed by MT users or developers to evaluate MT in real scenarios or by researchers in universities and other organisations, who have carried out detailed investigations into the efficacy of new methods or the improvability of existing ones;
b. are based on thorough research and/or sensible assumptions;
c. provide sufficient detail to be replicated and, in many cases, have been used subsequently by other experts in the field due to the reliability of their results.

Human evaluations of MT can be divided into two broad categories: (a) those which require access to the target text alone and (b) those which also rely on the source text or a human translation of the original, whether for the design of the evaluation, its implementation, or both. Types of evaluations within each of these categories are shown in Figure 3.1.

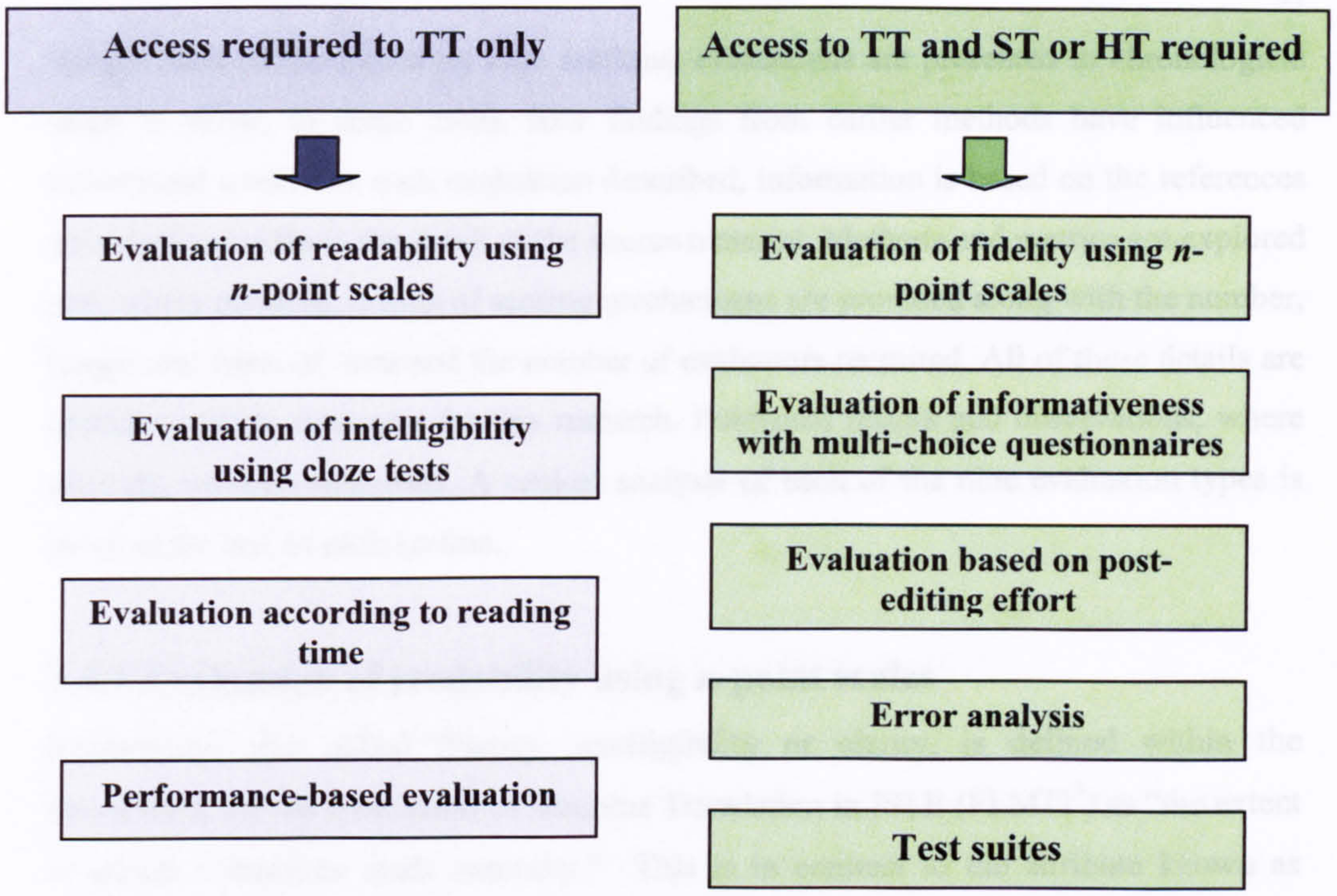


Figure 3.1: Types of human evaluations

Descriptions of evaluations in the first category are sub-divided as follows: Section 3.4.1 describes evaluations of readability using *n*-point scales to score text segments, and 3.4.2 explores the use of cloze tests to evaluate intelligibility. Evaluations based on the measurement of the time taken to read translated texts are presented in 3.4.3 and, finally, an example of performance-based evaluation is investigated in 3.4.4.

In the second category, Section 3.4.5 explores evaluations of fidelity using *n*-point scales, and 3.4.6 describes the use of multiple-choice questionnaires to evaluate the informativeness of MT output. Evaluations based on MT post-editing effort are presented in 3.4.7, and in 3.4.8 we consider methods involving the use of error classification schemes. Finally, in 3.4.9 we explore the use of test suites; although these tend to be used by MT developers to test the correct translation of isolated linguistic phenomena, an investigation of this form of evaluation may still prove useful for this research. After exploring the various ways in which MT output can be evaluated by humans, in Section 3.4.10 we arrive at a decision on which methods are most appropriate for this research.

Within each of the following nine sections, evaluations are presented in chronological order to show, in some cases, how findings from earlier methods have influenced subsequent work. For each evaluation described, information is based on the references provided and reflects the detail of the source material. Methods and metrics are explored and, where possible, details of scoring mechanisms are provided along with the number, length and types of texts and the number of evaluators recruited. All of these details are crucial to guide strategies for this research. Published results and observations, where relevant, are also described. A critical analysis of each of the nine evaluation types is given at the end of each section.

3.4.1 Evaluation of readability using *n*-point scales

Readability, also called fluency, intelligibility or clarity, is defined within the Framework for the Evaluation of Machine Translation in ISLE (FEMTI[3]) as "the extent to which a sentence reads naturally." This is in contrast to the attribute known as comprehensibility, defined on the same website as "the extent to which the text as a whole is easy to understand". The former attribute is of greater interest for this research, as its measurement provides more fine-grained results, it is the focus of a much larger number of published articles and, as a result, methods for its evaluation have improved over time.

The evaluation of readability using a scale to score individual sentences (or segments, such as headings) requires access only to the translations. Judges should be native speakers of the target language, but require no knowledge of the source language. In fact, some familiarity with the source language may influence a judge's opinion and, for this reason, evaluators with no such knowledge are preferred.

Assigning a score to a translated sentence is, of course, subjective and judgements can depend on a number of factors, including the level of appropriate knowledge or language proficiency of the evaluator, the length of the segment, the position of a sentence in the text as a whole (does it follow a series of 'good' sentences or 'bad' ones?), and human factors such as tiredness, boredom or hunger. In order to minimise the effects of subjectivity, readability evaluations tend to involve a large number of segments and texts, as the examples in this section will show, and each translation is

[3] http://www.isi.edu/natural-language/mteval/

evaluated by a number of subjects whose judgements are combined to produce a mean score.

Many articles have been published on the evaluation of readability and its related attributes by rating sentences on an *n*-point scale. These include methods described by Crook and Bishop, Leavitt, Pfafflin, Sinaiko, Van Slype and Vauquois in (Van Slype, 1979), (Maier et al., 2001), (Darwin, 2001), (Vanni and Miller, 2001) and (Babych et al., 2004). However, methods described below are examples of evaluations reported in more detail than the above.

3.4.1.1 Carroll: evaluation of intelligibility using a 9-point scale

Carroll's aim was to establish a standard procedure for measuring the quality of human and machine translations in the scientific domain. His experiments, described in (Pierce, 1966) involved the evaluation of intelligibility and informativeness, and the latter is described in 3.4.5. For the intelligibility evaluation, thirty-six random sentences were first selected from each of four Russian source texts. These were translated into English by three humans and three MT systems. Six sets of materials were then prepared, each containing one translation (from any system/human) of each of the 144 sentences so that no evaluator would see more than one translation of the same sentence.

Thirty-six native English speakers, half of whom had a good understanding of scientific Russian (the remainder having no knowledge of the source language) conducted the evaluation. Carroll's 9-point intelligibility scale, shown in Figure 3.2, was adapted from a psychometric technique (the method of equal-appearing intervals).

Each set of translations was evaluated by three monolingual and three bilingual subjects, providing six scores for each sentence, which were then combined to calculate a mean score. It was no surprise that the human translations scored consistently more highly than the MT. A high level of agreement was noted between scores given to the same sentence by individual evaluators, whether monolingual or bilingual. However, differences between scores were deemed significant enough to warrant at least three or four judges per sentence for future projects. It was also observed that bilinguals spent more time analysing the sentences, implying that monolinguals would be preferable as they work more quickly, and are likely to be easier to recruit.

9.	Perfectly clear and intelligible. Reads like ordinary text: has no stylistic infelicities
8.	Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or mildly unusual word usage that could, nevertheless, be easily "corrected"
7.	Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8
6.	The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Post-editing could leave this in nearly acceptable form
5.	The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible
4.	Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated
3.	Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesise the idea intended by the sentence
2.	Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical
1.	Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

Figure 3.2: Carroll's 9-point intelligibility scale

Carroll's findings still have value today, and a number of subsequent evaluations have incorporated his methods: Sinaiko in (Van Slype, 1979) modified Carroll's intelligibility scale to suit Vietnamese evaluators and used this for the evaluation of the English-Vietnamese LOGOS system. Similarly, Leavitt (Van Slype, 1979) proposed a method for measuring intelligibility adapted from Carroll's work. His method was designed for rating textual units (rather than isolated sentences, as in the case of Carroll) on a 9-point scale.

3.4.1.2 The Japanese government project for machine translation: evaluation of intelligibility using a 5-point scale

This evaluation, described in (Nagao et al., 1985), was conducted to test the feasibility of using MT to translate abstracts of scientific papers from Japanese into English.

Translations of 1,682 sentences from a scientific journal were evaluated for intelligibility and accuracy. The accuracy evaluation is described in 3.4.5.

One British and one American evaluator, each with experience of proof-reading and checking scientific translations, were used to represent average readers. They had no knowledge of Japanese. Untranslated words in the MT output were translated into English before the evaluation. The subjects then scored each sentence of the MT output using a 5-point scale, shown in Figure 3.3. Unlike Carroll's evaluation, lower numbers represent more intelligible sentences.

1. The meaning of the sentence is clear, and there are no questions. Grammar, word usage, and style are all appropriate, and no rewriting is needed.
2. The meaning of the sentence is clear, but there are some problems in grammar, word usage, and / or style, making the overall quality less than 1.
3. The basic thrust of the sentence is clear, but the evaluator is not sure of some detailed parts because of grammar and word usage problems. The problems cannot be resolved by any set procedure; the evaluator needs the assistance of a Japanese evaluator to clarify the meaning of those parts in the Japanese original.
4. The sentence contains many grammatical and word usage problems, and the evaluator can only guess at the meaning after careful study, if at all. The quickest solution will be a retranslation of the Japanese sentence because too many revisions would be needed.
5. The sentence cannot be understood at all. No amount of effort will produce any meaning.

Figure 3.3: The Japanese government project for machine translation: 5-point intelligibility scale

The authors remarked that many of the sentences were characteristically long and difficult, and it was suggested that the degree of difficulty of the source text should be taken into account when conducting any evaluation to make results more meaningful. It was also observed that it should be the users of MT output, whether specialists in a particular field or users of a technical manual, who determine the yardstick for intelligibility.

3.4.1.3 The DARPA series: evaluation of fluency using a 5-point scale

Between 1992 and 1994, the US Defence Advanced Research Projects Agency (DARPA) conducted a series of declarative evaluations as part of the Human Language Technology (HLT) initiative. The aim was to compare the various prototype systems it

was funding, along with a number of other commercial and institutional systems, to guide future research and development. Black box evaluations were required because the systems translated three different language pairs (French, Spanish and Japanese into English), and involved different approaches (statistics-based, knowledge-based and linguistic techniques).

The largest evaluation (White et al., 1994) resulted in a corpus of 100 general news stories (of approximately 400 words or 800 Japanese characters) in each of the three source languages, several machine translations of each source text and two expert human translations. Although a costly exercise, the carefully designed evaluation of such a large corpus provided reliable results, as it was based on methods that had evolved from lessons learned in previous evaluations in the series. DARPA translations and their scores have been widely used for subsequent research, particularly for the validation of scores produced by automated evaluation methods. Examples of this are provided in 3.5. Furthermore, the corpus has been valuable for this research and its use is described in 4.2.3.

The evaluation was carried out to measure the fluency, adequacy and informativeness of several machine translations and one human translation (as a control) of all 300 source texts. The adequacy and informativeness evaluations are described in 3.4.5 and 3.4.6 respectively. A number of evaluation books were compiled, each containing a mix of the three evaluation types with the aim of reducing fatigue among evaluators. Each book contained randomly selected translations from all source languages and systems and contained no more than one translation of any source text. Books were compiled so that texts appeared in different orders in each one. A practice text was provided for each evaluation type to give evaluators the opportunity to learn the task in an attempt to increase consistency. Evaluators took planned breaks to minimise omission errors, which had occurred in previous evaluations.

For the fluency evaluation, monolingual native speakers of English rated texts, sentence by sentence, using a simple 5-point scale, as shown in Figure 3.4. Systems were compared by calculating the average text score for each of the three attributes for each system. It was observed after the calculation of scores that twenty source texts rather than 100 would provide equally reliable results.

5 Excellent
4 Good
3 Fair
2 Poor
1 Very poor

Figure 3.4: DARPA 5-point fluency scale

3.4.1.4 TIDES: evaluation of fluency using a 5-point scale

The US Translingual Information Detection, Extraction and Summarization (TIDES) program is developing advanced language processing technology to enable English speakers to understand information in multiple languages. One of its ongoing objectives is to evaluate the performance of current DARPA-funded experimental MT systems (TIDES, 2005), again comparing them with human translations and commercial systems. Evaluations are conducted in the light of experience gained from the DARPA series described above, and translations are evaluated for both adequacy and fluency. Fluency is defined within the project as "the degree to which the translation is well-formed according to the grammar of the target language".

The focus is currently on the evaluation of Arabic and Chinese news stories translated into English. Each translated segment of a text is evaluated in sequence by at least two native speakers of the target language. No evaluator sees more than one translation of the same source text and the combination of texts for each evaluator is different.

An assessment user interface was developed as part of a purpose-built assessment system, enabling judges to carry out their work at a computer. For each sentence, the evaluator first provides a fluency judgement with access only to the translation. Having assigned a fluency score, the "gold standard" human translation appears, enabling the judge to evaluate adequacy (see 3.4.5). The fluency scale is shown in Figure 3.5. Evaluators are advised to spend no more than 30 seconds assessing each fragment for both fluency and adequacy, to provide intuitive judgements.

The assessment system, as well as automatically preparing and collating texts for evaluation, and assigning texts to judges, outputs scores for each segment along with

fragment and judge identifiers, ready for the computation of text and system scores for comparison.

How do you judge the fluency of this translation?

It is:

5 Flawless English

4 Good English

3 Non-native English

2 Disfluent English

1 Incomprehensible

Figure 3.5: TIDES 5-point fluency scale

3.4.1.5 Critical analysis

The evaluation of fluency using *n*-point scales has many advantages. Firstly, it can be carried out by monolingual judges with no knowledge of the source language or any special linguistic expertise. It is also resource-light, requiring access only to a set of segmented translations and a scoring metric. The evaluation involves relatively little preparation time, other than collating the evaluation packs, which should contain no more than one translation of any ST (to reduce the training effect). Texts should appear in a pseudo-random order, so that the same translation does not always appear, for example, at the beginning or end of the pack, as this may also have an effect on human scores. For large-scale evaluations it is advisable to automate the preparation of texts and collation of evaluator packs as in the TIDES project.

This type of evaluation is quicker to conduct than, for instance, evaluations of fidelity or informativeness (see 3.4.5 and 3.4.6.), which require subjects to read other material in addition to the MT output, such as the source text, a "gold standard" human translation, or multiple-choice questions. Furthermore, the method is system-independent and would be suitable for use with most text types and subject domains, provided that evaluators are representative of end-users.

Carroll's evaluation design was more complex than it needed to be, as individual sentences were extracted at random from the texts involved and re-collated to form

evaluator packs. This also meant that his sentences were rated out of context, which does not represent the actual end use of a translation, and this lack of text cohesion was bound to affect evaluator judgements. (For example, an incorrect anaphoric pronoun might not be penalised if a judge does not know its referent because it occurs within a different sentence.) The DARPA and TIDES evaluations, on the other hand, are simpler to design, in that texts are kept intact, and more reliable because sentences are judged in context.

It is important to note that the length of fluency scoring scales and their wordiness have reduced since Carroll's evaluation, reported in 1966. The earlier scales would have required more time to read and digest, and would make the evaluator's task more time-consuming. In contrast, the simple five-point scales used in the DARPA and TIDES projects are more easily understandable, and are based on experience gained from previous evaluations. A less wordy scale reduces the amount of time required for an evaluation, while increasing inter-evaluator agreement by not requiring judges to make unrealistically fine distinctions between different translated segments. However, much thought must be given to the choice of wording: the 'non-native English' and 'disfluent English' categories in the TIDES fluency scale are problematic, as they are open to wide interpretation.

The advantages of this method for evaluating readability outweigh the disadvantages, of which there are few. First, while the approach is suitable for end-users of MT systems, it provides little useful information for developers. For this research, however, scores from evaluators representing end users are important, as our aim is to develop an automated evaluation method that can predict these human intuitive judgements. Second, due to the subjective nature of this type of evaluation, a large number of sentences or texts are required and, preferably, more than one judgement per segment, making it more expensive and time-consuming than an objective evaluation requiring only one score per segment.

The readability evaluation method, using *n*-point scales to assign scores at sentence level, may be appropriate for this research. However, if MT systems are to be compared, it is crucial that no changes be made to the raw output. In (Nagao et al, 1985), translating all untranslated words in the MT output prior to the evaluation would have affected results by making the system appear better than it really was.

3.4.2 Evaluation of readability by the cloze test

The cloze procedure was originally devised by (Taylor, 1953) to evaluate the readability of human texts. This quality characteristic is assessed according to the ability of subjects to correctly guess words which have been omitted from the text, normally at regular intervals. The idea behind the cloze test is that a more intelligible text will be easier to complete correctly. Three experiments using the cloze test to evaluate MT output are described here.

3.4.2.1 Crook and Bishop

The idea of using the cloze test to evaluate MT was first suggested by (Crook and Bishop, 1965). Their experiment was later reported by (Halliday and Briss, 1977) and by Halliday in (Van Slype, 1979). Every eighth word in one human and one machine translation of the same source text was first blanked out. A number of readers were then given either the human or the machine translation and were asked to fill in the blanks with what they considered to be the missing words. Scores were calculated in two ways for each text based on (a) the number of answers comprising exactly the right word and (b) the number of correct answers including synonyms. A high correlation was observed between results obtained from the different readers.

3.4.2.2 Sinaiko and Klare

These two experiments, presented in (Sinaiko and Klare, 1972, 1973) are also described in (Somers and Wild, 2000). The cloze test was used to compare raw and post-edited MT output, produced by the English-Vietnamese LOGOS system, with a human translation. The evaluations differed from (Crook and Bishop, 1965) in that every fifth word was deleted and no score was awarded for synonyms. Two sets of scores were calculated: the percentage of correct responses and the percentage of answers omitted.

In the first evaluation, translations of three texts of 500 words were studied by 88 native speakers of English who completed the blanks in the original texts to provide a baseline for comparison, and 168 speakers of Vietnamese who evaluated the translations. In the second evaluation, just one text of 500 words was read by 57 English and 141 Vietnamese speakers. It was observed that the least technical text obtained higher scores. Results for the three translations proved to be statistically significant: even in the

second experiment, where the sample was much smaller, the human translation scored 55% accuracy, the post-edited MT scored 41% and the raw output scored 27%.

3.4.2.3 Somers and Wild

Experiments reported by (Somers and Wild, 2000) involved the evaluation of output from three MT systems and one human translation. The aims were (a) to determine whether results from the cloze test would rank the translations in the same order as subjective human judgements, and (b) to investigate the methodology of the cloze test for evaluating MT.

Three semi-technical French texts, each of around 500 words, were translated into English and words blanked out automatically. It was observed in a pilot study in which every fifth word was omitted (providing around 300 blanks in total) that subjects found the experiment too long and frustrating and that, particularly in the lower quality MT, it was extremely difficult to guess the missing words. As a result, every tenth word was blanked out for the full experiment. Twelve students with no knowledge of French were given one translation, each by a different system (or human), of each of the three source texts to complete, providing three different scores per translation. It was stressed that the evaluation was not a test of their ability, but of the scenario. They were not told that the texts were translations.

Only correct answers were given a score. Mean scores per system, normalised out of 50, ranged between 28.57 and 6.57. System scores were found to reflect initial subjective judgements about the three systems. Furthermore, differences between scores were, on the whole, found to be statistically significant, even though the sample size was only around 1,500 words. However, it was observed, particularly with lower quality MT, that certain factors such as the placement of gaps within the texts had a significant impact on scores.

The authors repeated the test using a more complex scoring system to determine how results and the ranking of systems might change if various wrong but plausible answers (eg. synonyms) were accepted and given half a mark. This more complicated method did not change overall results. It was also pointed out that the idea of acceptable synonyms introduces an element of subjectivity (ie. how does one decide which answers are acceptable?). This is not the case with the simpler method, where only correct

answers are given a score. It was concluded that the cloze test appeared to be a good indicator of relative translation readability, but that tests on a much larger sample would be needed to confirm this.

3.4.2.4 Critical analysis

This second approach to assessing readability has many of the same advantages as the evaluation of fluency using *n*-point scales. Evaluators can be monolingual judges with no knowledge of the source language and the method is resource-light, requiring only translations with blanked out words and a set of clear instructions. Furthermore, the preparation of the evaluation material is easily automatable. Again, this type of evaluation is much quicker to carry out than evaluations of fidelity, as only the target text is read by evaluators. The method can be used with different text types and subject domains, provided that evaluators know the subject matter. In fact, when working with technical language, domain-specific expertise is essential to achieve reliable results, as a lack of specialised knowledge can significantly reduce the number of correct answers given. For this reason, the method is more suited to the evaluation of general texts.

This type of evaluation is objective, but due to various influencing human factors (eg. subject knowledge, language proficiency, levels of frustration with the experiment), acquiring scores from several judges for each translation is advisable. Three readers per text, as in (Somers and Wild, 2000) seem sufficient, as results based on this number of subjects were deemed to adequately predict intuitive judgements. Further, even with a small sample size, differences between system scores proved to be statistically significant.

A major disadvantage of this method is that results have proven less reliable when comparing systems with lower quality output. It is far more difficult to guess the omitted words in poorer translations, which makes suitably fine-grained comparisons between these systems more difficult.

A further issue is the placement of blanked out words: in order to be objective, intervals between blanks should be equal. However, this means that in two translations of the same source text, the gaps will obviously coincide with different words, some of which will be easier to complete than others. For example, a large number of omitted function words in one translation may give that text an unfair advantage over one with a greater

number of blanked out content words. Furthermore, a text containing many dates, numbers or named-entities, for example, will be more difficult to complete than a text with a lower frequency of these kinds of tokens. One can, of course, select particular tokens to blank out in the translations. (Miller, 2000), for example, used the cloze test to evaluate the translation of prepositions in MT output. Adapting the procedure in this way provides information about specific errors made by systems, the number of which may or may not reflect intuitive quality judgements of the translations.

It should be noted that even when subjects see each translation only once, evidence of the training effect has been noted. In a cloze test evaluation described by (Miller et al., 2001), subjects pointed out that they were able to learn some untranslated Spanish words in the texts as they progressed through the translations, which artificially improved comprehensibility and, therefore, the subjects' ability to produce correct answers. However, there was no mention of this effect occurring among subjects reading translations from Chinese, indicating how the source language can affect results.

Finally, while this evaluation method may be appropriate for end-users, since it has been shown to reliably predict intuitive judgements of readability, it provides little useful information for developers. Moreover, the method does not produce individual scores at sentence (or segment) level, making results insufficiently fine-grained for many purposes.

3.4.3 Evaluation of readability according to reading time

This type of evaluation involves the measurement of time required to read and understand a text, or to realize its unintelligibility, but not to memorise it (Van Slype, 1979). This is based on the assumption that a poorer quality translation will take longer to attempt to understand. Relatively few evaluations have been based on the measurement of reading time; however some are briefly mentioned in (Van Slype, 1979) and are summarised below.

3.4.3.1 Examples of evaluations by reading time

Van Slype states that reading time can be assessed in various ways and cites the following examples:

- Dostert suggests measuring the percentage of additional time required by native speakers of the target language to read a machine translation as opposed to an original text. We assume that this would be a comparable text of the same length.
- Carroll proposes measuring the time required by an evaluator to read each sentence in a sample.
- Van Slype suggests the same method, but measuring the time required to read the whole text.
- Pfafflin and Orr propose measuring the time taken to answer multiple-choice questions on the MT output.
- Sinaiko suggests calculating the time taken to complete a cloze test.

3.4.3.2 Reeder: Evaluation by time taken to distinguish MT from HT

This intelligibility evaluation by (Reeder, 2001 and 2004) was derived from a second language acquisition experiment, showing that native-speaker essays can be distinguished from non-native essays by reading fewer than 100 words. The aims were to ascertain whether this technique could be applied to the evaluation of MT output and to acquire information on the kinds of errors encountered that led subjects to identify a text as MT. Although the objectives differ from methods in 3.4.3.1, the number of words an evaluator needed to read before identifying a text as HT or MT was found to reflect human fluency judgements.

Reeder used Spanish-English machine translations from five systems along with one human translation of fifty texts from the DARPA corpus. Headlines, which are characteristically difficult to translate by MT for a number of reasons, were removed to prevent subjects from distinguishing MT by looking no further than the heading. The first portion of each translation (up to 140 words) was used for the experiment.

Fifty native speakers of English were each given six extracts of different human and machine translations to read, providing one judgement per text. They were asked to read up to the point at which they felt that they could decide that a text was produced by a human or a machine. A maximum of three minutes per text was permitted. Subjects marked the word at which they made their decision.

An analysis of results showed that across all texts, correct distinctions were made 87.7% of the time. Importantly, a correlation was found between the number of words the

evaluator needed to read before making a decision and DARPA fluency scores by system. Essentially, the more words an evaluator needs to read, the more intelligible (and human-like) the translation is. Particular error types were found to be good indicators of MT. The most frequent were incorrect prepositions and word ordering.

3.4.3.3 Critical analysis

Like methods described in 3.4.1 and 3.4.2, the above evaluations are simple to prepare, requiring only the target texts, and are relatively quick to conduct. Furthermore, monolingual judges with no knowledge of the source language can be used and methods are appropriate for most text types and domains, provided that evaluators have the appropriate subject knowledge.

The main problem with this kind of evaluation is that it is more suitable for providing evaluation results at the text level, which are insufficiently fine-grained for much MT research. Carroll's suggestion of measuring the time required to read each sentence is also problematic. This involves the difficult task of measuring very short time spans, and accuracy might be called into question, particularly for very short sentences. Measuring the time taken to complete other kinds of evaluations (Pfafflin and Orr, Sinaiko in (Van Slype, 1979)) involves a greater amount of work, as it relies on the preparation of other experiments against which results are compared.

A further problem concerns the extent of subjectivity linked to this kind of experiment. At what point does a judge decide whether he/she has understood a sentence? At what point does an evaluator give up trying to make sense of 'word salad'? Furthermore, if we go back to Van Slype's definition of this type of evaluation at the beginning of 3.4.3, it becomes clear that the length of time required to read and understand a very good translation or to realize the unintelligibility of a very bad one is likely to be similar (ie. relatively short); a greater amount of time is likely to be spent reading translations whose quality lies somewhere in the middle. In addition, some people read more quickly than others, some will try harder and spend more time trying to make sense of something. With Reeder's method, language proficiency and knowledge of the kinds of errors that occur in human or machine translations will certainly affect results. It is not yet clear how many texts would be required to minimise these effects and produce optimum results. This also begs the question: is one judgement per text sufficient?

The above methods are more appropriate for end-users of MT systems than for developers, but Reeder's approach additionally provides useful information for research and development in terms of the kinds of errors found in MT that distinguish it from HT.

3.4.4 Performance-based evaluation

Performance-based evaluation has been used to assess the quality of MT output according to the ability of subjects to follow practical instructions in a translation. One example of this kind of evaluation is described below.[4]

3.4.4.1 Sinaiko: Measurement of usefulness by performance

The measure of translation quality for this approach combines elements of fidelity and readability, although this evaluation reported in (Van Slype, 1979) involves access only to the MT output. The method was used to evaluate MT from the English-Vietnamese LOGOS system.

The sample of texts for evaluation comprised translated extracts from maintenance manuals, containing lists of instructions for carrying out particular processes. Native speakers of the target language used the MT output to attempt to perform the tasks according to the instructions. The system was evaluated by measuring mistakes in performance for each instruction by calculating the number of tasks performed with (a) no errors, (b) a minor error and (c) a major error.

3.4.4.2 Critical analysis

In comparison with the other methods described above, this evaluation is far more complex. Although it still has the advantage of requiring monolingual subjects who need access only to the MT output, the preparation of particular equipment is needed for evaluators to carry out the required tasks.

Sinaiko claimed that his method was objective and effective and took into account many aspects of translation quality. However, the scoring metric is problematic. First,

---

[4] Work related to the MT Functional Proficiency Scale project reported by (White and Taylor, 1998), (Taylor and White, 1998) and (White et al., 2000) is not explored here, as the aim was to develop a test suite of errors for predicting the text-handling tasks that MT output from different systems would support.

decisions on what constitutes a major and minor error are subjective and second, the ability to complete one instruction in a list must, to a greater or lesser extent, be influenced by the quality of previously translated or subsequent instructions in the text. Furthermore, the method would be unable to provide fine distinctions between low quality machine translations.

The author recognised that the method was slow and expensive. We assume that this was due to the fact that tasks are time-consuming to complete and that equipment is required in order for subjects to follow the instructions. Furthermore, the evaluation is restricted to only a small number of text types, making it less flexible than other kinds of evaluation.

3.4.5 Evaluation of fidelity using *n*-point scales

In his Critical Report, (Van Slype, 1979) defines fidelity as the "subjective evaluation of the degree to which the information contained in the sentence of the original text reappears without distortion in the translation". In the same report, Halliday describes fidelity as the "measurement of the correctness of the information transferred from the source language to the target language". Measures of fidelity include informativeness, accuracy and adequacy, terms which are used in some of the evaluations described below.

(Van Slype, 1979) emphasises the difficulties involved in the evaluation of fidelity, because "each sentence conveys not a single item of information or a series of elementary items of information, but rather a portion of message or a series of complex messages whose relative importance in the sentence is not easy to appreciate." The evaluation of fidelity is subjective, as the various pieces of information in a sentence can differ in importance for each reader.

The methods described here represent a small number of the many published fidelity evaluations using n-point scales. Other work includes methods by Crook and Bishop, Leavitt, Miller and Beebe-Center and Van Slype in (Van Slype, 1979), (Jordan et al., 1993), (Bohan et al., 2000), (Miller et al., 2001), (Darwin, 2001) and (Babych et al., 2004a, 2004b). The technique has evolved over time, as can be seen from the experiments reported below. Today, the method involves either monolingual evaluators who use a human "reference translation" for comparison, or bilingual subjects who

compare the MT output with the source text. Scoring descriptions, segment lengths, the number of texts and evaluators tend to differ from one evaluation to another.

3.4.5.1 Carroll: evaluation of informativeness using a 10-pt scale

This experiment accompanied Carroll's evaluation of intelligibility (Pierce, 1966) described in 3.4.1. The same sets of materials were used, containing 144 sentences out of context from scientific texts, translated from Russian into English by one human and three MT systems. The same evaluators - eighteen monolingual English speakers and eighteen native English speakers with a good understanding of scientific Russian - conducted the evaluation. Due to the subjective nature of the method, three monolingual and three bilingual evaluators separately rated each set of materials, sentence by sentence, providing six scores per text. No evaluator saw more than one translation of the same sentence.

The informativeness evaluation was designed in a slightly different way for the two sets of judges. Bilingual subjects read each translated sentence and then rated the *original* Russian sentence for its informativeness compared with the translation using the 10-point scale in Figure 3.6. Monolingual evaluators compared each translated sentence with an expert human translation, which was then rated for its informativeness using the same scale. Lower scores indicate a more faithful translation. The scale was adapted from the same psychometric technique as that used for the intelligibility evaluation.

The mean sentence score per system/human was calculated to provide the final results. As predicted, scores for human translations were consistently higher. An important finding was that a high correlation was observed between scores from the intelligibility and informativeness evaluations. Furthermore, a strong correlation was noted between scores from the monolingual and bilingual evaluators (just as in the intelligibility evaluation) and bilingual judges were found to spend more time reading and rating the sentences.

9.	Extremely informative. Makes "all the difference in the world" in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation)
8.	Very informative. Contributes a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the reader's impression of the meaning intended, although not so much as to change or reverse the meaning completely
7.	(Between 6 and 8)
6.	Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader "on the right track" as to the meaning intended
5.	(Between 4 and 6)
4.	In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words
3.	By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different "twist" to the meaning conveyed by the translation. It adds no new information about sentence structure, however
2.	No really new meaning is added by the original, either at the world level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended
1.	Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced
0.	The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.

Figure 3.6: Carroll's 10-point informativeness scale

3.4.5.2 The Japanese government project for machine translation: evaluation of accuracy using a 7-point scale

This experiment was conducted in conjunction with the evaluation of intelligibility described in 3.4.1 and is also reported in (Nagao et al., 1985). The same 1,682 sentences from scientific papers were used, translated from Japanese into English. Four bilingual judges evaluated how much of the meaning of the original text was conveyed in the MT output. Accuracy was evaluated on a 7-point scale shown in Figure 3.7.

In support of Carroll's findings results showed that in most cases the lower the accuracy score, the lower the intelligibility score. Due to the characteristically long and difficult sentences, it was suggested that the difficulty of the source text should be taken into consideration to make results more meaningful in subsequent experiments.

0. The content of the input sentence is faithfully conveyed to the output sentence. The translated sentence is clear to a native speaker and no rewriting is needed.
1. The content of the input sentence is faithfully conveyed to the output sentence, and can be clearly understood by a native speaker, but some rewriting is needed. The sentence can be corrected by a native speaking rewriter without referring to the original text. No Japanese language assistance is required.
2. The content of the input sentence is faithfully conveyed to the output sentence, but some changes are needed in word order.
3. While the content of the input sentence is generally conveyed faithfully to the output sentence, there are some problems with things like relationships, between phrases and expressions, and with tense, voice, plurals, and the positions of adverbs. There is some duplication of nouns in the sentence.
4. The content of the input sentence is not adequately conveyed to the output sentence. Some expressions are missing, and there are problems with the relationships between clauses, between phrases and clauses, or between sentence elements.
5. The content of the input sentence is not conveyed to the output sentence. Clauses and phrases are missing.
6. The content of the input sentence is not conveyed at all. The output is not a proper sentence; subjects and predicates are missing. In noun phrases, the main noun (the noun positioned last in the Japanese) is missing, or a clause or phrase acting as a verb and modifying a noun is missing.

Figure 3.7: The Japanese government project for machine translation: 7-point accuracy scale

3.4.5.3 The DARPA series: evaluation of adequacy using a 5-point scale

The DARPA adequacy evaluations were carried out at the same as the fluency evaluations described in 3.4.1.3, where background information on the project is also provided. The same news texts were evaluated, translated into English by various MT systems and two humans, from French, Spanish and Japanese.

As described in 3.4.1.3, 100 native English speakers each completed one evaluation book, containing a mix of all three evaluation types involved in the experiment. No evaluator saw more than one translation of the same text and each translation was rated

by a different evaluator for each of the three attributes. A practice text was provided so that evaluators could learn the task in an attempt to increase consistency.

For the adequacy evaluation, candidate translations (whether machine or human) were divided into fragments (usually shorter than a sentence) and placed alongside their corresponding reference translations (expert translations, conveying all the factual information of the original without any stylistic flourishes). Evaluators compared each fragment in the candidate translation with the reference, and scored each one using a 5-point scale, according to how much of the original content had been preserved in the translation, regardless of imperfect English. The scale is shown in Figure 3.8.

5	All meaning expressed in the source fragment appears in the translation fragment
4	Most of the source fragment meaning is expressed in the translation fragment
3	Much of the source fragment meaning is expressed in the translation fragment
2	Little of the source fragment meaning is expressed in the translation fragment
1	None of the meaning expressed in the source fragment is expressed in the translation fragment

Figure 3.8: DARPA 5-point adequacy scale

Scores for each fragment were combined to produce a mean text score for each attribute for each system. A correlation was again observed between scores for fluency and adequacy, although this found to be stronger for the higher and lower values.

3.4.5.4 TIDES: evaluation of adequacy using a 5-point scale

The current TIDES project involves the evaluation of both fluency and adequacy. Background information and the fluency evaluation method are described in 3.4.1.4.

Adequacy is defined as "the degree to which the translation communicates information present in the original" (TIDES, 2005). Evaluators compare each text fragment (translated from either Arabic or Chinese into English) with a "gold standard" translation, selected by a bilingual linguist and a senior annotator from a set of human translations. Each fragment is separately evaluated by at least two native speakers of the target language. No evaluator sees more than one translation of the same source text and the combination of texts for each evaluator is different.

For each fragment, subjects first provide a fluency judgement before the gold standard translation appears on the screen, enabling judges to evaluate adequacy. The next fragment in the news story is then presented. The adequacy scale, which is much the same as the one used in the 1994 DARPA evaluations, is shown in Figure 3.9.

How much of the meaning expressed in the gold standard translation is also expressed in the target translation?

5 All

4 Most

3 Much

2 Little

1 None

Figure 3.9: TIDES 5-point adequacy scale

3.4.5.5 Critical analysis

This kind of evaluation can be carried out by monolingual or bilingual judges and, as Carroll discovered, a strong correlation was observed between scores from both sets of judges. Where practical, the preparation and collation of evaluator packs can be automated, ensuring that all translations receive the same number of judgements, that texts are placed in a pseudo-random order in each pack and that no set contains more than one translation of each original. The method can be used with most text types and subject domains, but when evaluating texts in specialised domains, evaluators must be familiar with the subject matter.

The main disadvantage of this method, when compared with readability evaluations, is that additional material must be prepared and analysed, making the whole process more expensive and time-consuming. While source texts are readily available for comparison, bilingual judges must be found, and when using monolingual subjects who are easier to recruit, translators must be employed to produce gold standard translations.

The approach is suitable for end-users to compare output from different MT systems, but provides little useful information for developers, eg. in terms of the kinds of errors

that systems produce. Further, due to the element of subjectivity, a large sample is required and, preferably, at least two judgements per segment, making it more expensive and time-consuming than more objective evaluation methods.

When comparing the evaluations described in this section, the more recent DARPA and TIDES methods are preferable in that (a) the translations remain intact for evaluation, so that context can be taken into consideration and (b) the scoring scales are more concise than those by Carroll (Pierce, 1966) and (Nagao et al., 1985) and are, therefore, easier to understand and quicker to use. Furthermore, the accuracy scale used in the Japanese government project requires some specialist linguistic knowledge to be understood and is suitable in its present form only for evaluating MT, as it contains some criteria which are not applicable to human translations. As with the readability scales presented in 3.4.1, the simpler scales used by DARPA and TIDES can only reduce the amount of time required for an evaluation, while increasing inter-evaluator agreement by not requiring judges to make too fine distinctions. It is also worth noting that Carroll's questionable idea of scoring the original text (or human translation) rather than the MT output has not been adopted since, and subsequent evaluations have focused on the scoring of the translation itself.

Finally, correlations observed between intelligibility and informativeness scores indicate that the simpler evaluation for intelligibility (involving only the translated sentences, and requiring less reading time) may be sufficient for some purposes. The correlation between scores for these attributes will be investigated as part of this research in order to support our hypothesis that an automated evaluation method based on access to the MT output alone can predict judgements of the quality of a translation as a whole.

3.4.6 Evaluation of informativeness by questionnaire

The aim of this kind of evaluation is to test whether enough correct information is conveyed in the translation to enable evaluators to answer questions about its content. Although evaluators do not see the source text or a gold standard translation, the method relies on access to the original (or a human translation) for the preparation of multiple-choice questions on the content of the text. Two such evaluations are reported here. Similar approaches devised by Leavitt and Orr are reported in (Van Slype, 1979).

3.4.6.1 Sinaiko: Measurement of comprehension by the knowledge test

Sinaiko used this method to evaluate output from the English-Vietnamese LOGOS system and recommended it for further use (Van Slype, 1979). A questionnaire was written in both the source and target language, based on information contained in the original text. This contained questions intended to assess the knowledge that a reader can gain from a candidate text. Two groups of readers who were speakers of the source or target language then took the appropriate questionnaire and attempted to answer the questions based on the content of the translation or the original text. The mean number of correct answers per group was then calculated in order to make a comparison.

3.4.6.2 The DARPA series: evaluation of informativeness using a multiple-choice questionnaire

This was the third evaluation method used as part of the 1994 DARPA series, along with assessments of fluency and adequacy. As mentioned in 3.4.1.3 and 3.4.5.3, the three evaluation types were combined in carefully designed evaluator booklets (to reduce boredom and fatigue) and a practice text was provided to enable evaluators to familiarise themselves with the task prior to the evaluation. All texts involved in the experiment (one human and various machine translations of 100 French, Spanish and Japanese news articles) were evaluated for informativeness using this method.

Each translation, of approximately 400 words, was accompanied by six multiple-choice questions on its content. Evaluators chose from six possible answers to each question. Systems were compared by calculating the average text score for each system. Results showed a strong correlation between scores for adequacy and informativeness. This is not surprising, as both are measures of fidelity.

3.4.6.3 Critical analysis

This evaluation requires only monolingual subjects and, although Sinaiko's method requires speakers of both the source and target languages, the DARPA evaluation found target language only speakers sufficient. This approach can be used with most text types provided that suitable questions can be devised on the content.

There are two main problems with this type of evaluation. The first concerns the preparation of questions. Each original text (or gold standard translation if available)

must first be read and understood before questions can be devised. Content must be selected on which to base the questions, which must then be carefully written. For multiple-choice questionnaires, a set of reasonable alternative answers must also be devised. This all makes for a time-consuming and expensive process, particularly when evaluating a large number of texts, and one that cannot be automated.

The second problem concerns the granularity of the scoring method. In the DARPA evaluation, a score of between 0 and 6 is calculated for each text of around 400 words, and this begs the question; can six facts convey the quality of a whole text? Such a limited scoring mechanism is inadequately fine-grained for many needs, and particularly when resources are not available to replicate such a large scale evaluation that produces a large number of scores per system.

Sinaiko claims that this method is objective, in that there are only correct or incorrect answers to the questions provided. However, the choice and number of questions and an evaluator's knowledge of a given subject domain can affect results, as can the varying ability of evaluators to perform the task. This means that more than one set of answers per text is required. Furthermore, the whole text should be read (and some sections more than once) in order for subjects to best answer the questions, making the evaluation relatively time-consuming.

The strong correlation between scores for adequacy and informativeness in the DARPA evaluations indicates that just one method would suffice; bearing in mind the amount of time required to prepare and conduct the two evaluations, the adequacy evaluation seems preferable as the scores it produces are more fine-grained.

3.4.7 Evaluation by post-editing effort

According to Allen in (Somers, 2003), "the task of the post-editor is to "edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)". Since the main reason for using MT is to increase productivity, output can be assessed according to (a) the time required to post-edit a given sample of raw MT to an agreed standard or (b) the number of corrections needed. Any company using this method to compare systems must, of course, ensure that texts reflect their intended use of MT.

(Krings, 2001) believes that "post-editing effort is the key issue in the evaluation of the practicality of machine translation systems" and suggests that it can be measured on three levels; temporal, technical or cognitive. The temporal level concerns the time taken to post-edit a given sample of MT output; the technical level involves counting the number of deletions, insertions and instances of word reordering, and the cognitive level concerns the investigation of processes used to remedy deficiencies in the target text, which might involve recording the post-editor's thoughts using Think Aloud Protocols (TAPs). All three methods could, in some way, be measured automatically. For instance, a tool called Translog was developed at the Copenhagen Business School for monitoring keyboard activity during translation. (Hansen, 1999) used this to explore correlations between post-editing effort and source text translatability, measuring (1) the time taken to post-edit a translation, (2) the number of words inserted and deleted, cut and paste actions and dictionary look-ups and (3) the number and duration of pauses (in an attempt to quantify cognitive effort).

Methods for evaluating MT according to post-editing effort are described by Andreewsky, Chaumier, Dehaven, Hofstetter and Van Slype in (Van Slype, 1979), but the information provided lacks detail. (Roudaud et al., 1993) describe their preparatory work for using post-editing time to evaluate the French-English ARIANE system, but no results are available. The two evaluations described below, both focusing on correction time, offer more detail on this type of evaluation, and present some useful findings.

3.4.7.1 Wagner: measurement of post-editing time

This method was one of four approaches explored by (Wagner, 1998) to compare output from four English-German MT systems. For the measurement of post-editing time, each system translated three source texts containing a total of around 900 words. Samples of MT from all four systems were first combined to create a test piece to be corrected with access to the source text by all sixteen subjects (native German speakers who were advanced students of English) to provide a measure of their pace and style. Each student then revised one translation, providing four measurements for each system. Subjects were instructed to correct grammatical and lexical errors but not the style. The time taken to post-edit the standard text, the candidate translation and the average ratio time per system were calculated.

Wagner concluded that while this is a very practical MT evaluation method, it has its drawbacks. For example, she found that individual working styles and paces vary considerably; subjects took between 18 and 61 minutes to complete the test piece. For the evaluation itself, the quickest post-editor took 40 minutes and the slowest 133 minutes. Even the ratios between the time taken by each subject to post-edit the test piece and the evaluation text varied, and an insufficient correlation between post-editors meant that systems could not be ranked reliably. Some of the students admitted that they were not sure if they had been consistent in their level of correction in the test piece and the evaluation text; some were tempted to retranslate sentences because it was easier than post-editing the output when the quality was particularly bad. Furthermore, post-editing time was also found to be influenced by personal factors, such as knowledge of the terminology, working style, distractions and the ability to concentrate.

3.4.7.2 Minnis: evaluation based on post-editing

This experiment (Minnis, 1993, 1994) was devised to test the effectiveness of a method for evaluating output from the same MT system over time, assuming that the amount of post-editing required would reduce with an improved version of the system. The method was designed to be potentially useful for users and development managers to (a) quantify improvements in system output and (b) show that investment in a system is worthwhile.

A technical manual containing 859 sentences was translated from Japanese into English by two consecutive versions of the same MT system. All unknown words were entered into the system dictionaries prior to the evaluation. Three monolingual and three bilingual subjects were recruited for the experiments. The evaluation comprised four stages:

1. After initial training, monolingual subjects assigned a pass/fail mark for understandability to each sentence in the raw MT output.
2. Monolinguals post-edited the MT and recorded the time spent on each sentence. Post-editing time was defined as "the time in seconds for a monolingual to type from scratch a complete and understandable version (in their opinion) of a given sentence, under standard operating conditions." This was intended as a standardisation procedure, since all subjects were meant to be roughly equal in

ability. Monolinguals were chosen as they were cheaper to employ and brought no semantic knowledge about the source language to bear on the rating or post-editing.

3. Bilinguals assigned a pass/fail mark for accuracy to the post-edited sentences, a pass being assigned to all sentences that faithfully conveyed the meaning of the source text. (This was based on the assumption that it is easier to assign a pass/fail mark than a value from a scale.) This was intended to test the system's ability to provide output which, when tidied up by a monolingual, could provide an accurate translation.
4. Accuracy pass/fail judgements were confirmed by other monolingual and bilingual subjects to reduce subjectivity.

Results showed that the number of failed sentences fell from 41% to 28% in output from the updated version of the system. The average post-editing time decreased from 5.5 hours to 4.7 hours, and some correlation was found between results for understandability and post-editing time. Some evidence of the training effect was observed, as subjects began to learn consistently incorrect structures which helped them to post-edit later sentences more quickly or in a better way.

3.4.7.3 Critical analysis

Both methods described above represent very practical ways in which to evaluate MT output, particularly for businesses or institutions that work with or are considering using MT. The time required to revise raw output can be used to compare different systems or to calculate possible savings by using MT rather than translating from scratch. However, as these two quite different approaches indicate, there are various drawbacks involved.

Wagner's method requires bilingual subjects and, ideally, these should be experienced translators or post-editors rather than language students. The former are used to correcting the same kinds of repeated errors, which typically occur in MT output and are less likely to become annoyed with the process than non-specialists. Experienced post-editors develop strategies for correcting repeated errors as quickly as possible, such as using macros or performing global search and replace actions. As a result, specific instructions should be provided to ensure that all subjects adopt (or refrain from using) the same strategies to make their correction times more comparable. These instructions

are dependent on whether revision time is to be measured at the sentence or the text level.

Minnis' decision to use monolingual subjects to post-edit the output after dictionary update makes for quite a different experiment, which tells us more about the understandability of MT for monolingual users. Minnis attempts to make correction times more comparable by instructing subjects to retype each sentence but, nevertheless, typing speed and the ability to use a keyboard will vary. The evaluation still requires the expense of recruiting bilingual judges to rate the post-edited output and there is no assessment of lexical coverage.

This method of evaluation can be expensive because a large number of post-editors and translated sentences are required to produce reliable results, due to a host of different factors, including variations in working pace, attitudes to the task and perceptions of acceptable quality. For these reasons, a good deal of preparation and evaluator training (eg. by use of a test piece, as in Wagner's experiment) is required, and specific guidelines must be provided on the level of post-editing required.

3.4.8 Evaluation by error analysis

Error analysis is essential for developers who often use test suites containing instances of particular grammatical phenomena to evaluate systems with access to internal components. However, a number of black box evaluations have also been reported, involving the use of error classification schemes to evaluate MT from a user's point of view. With access to the source text, errors are systematically classified and counted, with or without weighting, to provide information on the number of errors in any given output, enabling users to compare different systems.

A number of articles on the development and use of error classifications have been published, some in greater detail than others. For example, error schemes by Chaumier and Green (both for English-French), Knowles (Russian-English) and Masterman (no language pair stated) are described in (Van Slype, 1979), but no results are reported. (Loffler-Laurin, 1983) presents a typology of errors for MT, but rather than being used for evaluation, her study is intended to help developers to make system improvements and to determine which kinds of documents are suitable for MT. (Roudaud et al., 1993) present a detailed French-English error classification for use by post-editors, but only

initial preparatory work is described. (Bohan et al., 2000) developed a scheme containing 151 categories to help developers identify areas for system improvement. However, only a short extract from the scheme is provided and, due to its length and complexity, the method requires a lot of training as evaluators are likely to misinterpret the metrics and/or instructions. (Correa, 2003) presented a typology of errors common in English-Spanish MT output to provide information to system developers. His scheme is not suitable for users in its current state, as some of the categories are descriptive of the sources of error, such as segmentation, source analysis and target grammar problems. Finally, (Guessoum and Zantout, 2005) developed a non-exhaustive English-Arabic scheme to manually evaluate lexical and grammatical coverage, semantic correctness and correct pronoun resolution.

The two evaluations described below provide details of the entire error classification schemes used and their development and implementation to compare different MT systems. Both experiments present useful observations and findings.

3.4.8.1 Flanagan: error classification

This evaluation, reported in (Flanagan, 1994), involved the design of an error classification scheme for use by potential users of MT to compare different systems. As such, the focus is on the identification of errors in the output rather than their causes. A classification scheme was developed by identifying the most frequent errors in French MT output translated from English sentences in the Hewlett-Packard test suite. The scheme is shown in Table 3.2.

The author observed that, although some error categories can apply to many languages, a unique error classification is required for each language pair according to the types of errors that occur. For instance, when the English-French scheme was adapted for English-German evaluation, the 'Elision' category was removed, as it was not applicable, and three new categories were added: 'Relative pronoun absent or incorrect', 'Case ending incorrect' and 'Punctuation incorrect, absent or unneeded'. Flanagan also found that the impact of particular error categories can differ with different language pairs, and some cause additional errors.

Category	Description
Spelling	Misspelled word
Not found word	Word not in dictionary
Accent	Incorrect accent
Capitalization	Incorrect upper or lower case
Elision	Illegal elision or elision not made
Verb inflection	Incorrectly formed verb, or wrong tense
Noun inflection	Incorrectly formed noun
Other inflection	Incorrectly formed adjective or adverb
Rearrangement	Sentence elements ordered incorrectly
Category	Category error (eg. noun vs verb)
Pronoun	Wrong, absent or unneeded pronoun
Article	Absent or unneeded article
Preposition	Incorrect, absent or unneeded preposition
Negative	Negative particles not properly placed or absent
Conjunction	Failure to reconstruct parallel constituents after conjunction, or failure to identify boundaries of conjoined units
Agreement	Incorrect agreement between subject-verb, noun-adjective, past participle agreement with preceding direct object, etc.
Clause boundary	Failure to identify clause boundary, or clause boundary unnecessarily added
Word selection	Word selection error (single word)
Expression	Incorrect translation of multi-word expression

Table 3.2: Flanagan's error classification

A test suite of 910 sentences was translated into French and German by three MT systems. Bilingual subjects used the error classification to annotate errors with access to the source text. The classification process took 160 person hours in addition to training costs.

Flanagan suggested that when using the classification scheme for evaluation, error types should be ranked in terms of importance for the user, as simply counting the number of errors in a translation is not meaningful (some involve only one word, some involve phrases or word ordering and one error can lead to another). For this evaluation, error types were ranked in terms of their effect on (1) improvability and (2) intelligibility. For instance, 'Not found words' are easily improvable as they can be added to system dictionaries, whereas 'Clause boundary' and 'Conjunction' errors are among the least improvable errors, requiring analysis and improvement by developers. In terms of

intelligibility, 'Elision', 'Accent' and 'Capitalization' errors have little impact, whereas 'Word selection' errors can have a far greater effect.

The author provides results for the English-French evaluation, in terms of the number of each kind of error identified in the output from each system. These error totals provided sufficiently fine distinctions between output quality to be able to reliably compare systems, and the total number of some individual error types reflected the same ranking of systems.

Flanagan concludes that because MT evaluation is subjective, using an error classification scheme enables judgements to be more consistent. She points out that an error scheme is simple to design and flexible, because categories can be added or deleted and ranked according to user needs.

3.4.8.2 Wagner: error classification

This evaluation (Wagner, 1998) was carried out as part of an experiment to explore four different evaluation methods suitable for end-users. (Her measurement of post-editing time is reported in Section 3.4.7.1.) Raw output from four English-German MT systems was evaluated with no prior dictionary updating, as possibilities for updating differed considerably across systems. The source text comprised 596 words (27 sentences) from a magazine editorial. The error classification scheme, shown in Figure 3.10, was based mostly on categories devised by (Roudaud et al., 1993).

Wrongly translated word or expression
Incorrect order of words
Addition or removal of words
Untranslated word
Wrong punctuation
Typographical problems
Truncated word
Stylistically unhappy choice of words

Figure 3.10: Wagner's error classification

In order to assign the error categories, the author minimally corrected each sentence to an 'acceptable' standard, aiming to stay as close as possible to the raw output. Errors were then counted, but not weighted. The number of errors per system ranged between 105 and 197.

Wagner concluded that this method has an advantage over the use of test suites, in that real material can be used, containing long and complex sentences and various forms of punctuation. However, the author noted that constant, time-consuming comparisons had to be made between the different translations to ensure consistency when assigning the error types. Furthermore, on some occasions more than one category could be assigned, and decisions had to be taken on the number of errors to record. The author did not see this as a problem, provided that the annotator remains consistent in the method of error assignment. The time taken to annotate all errors in the 27 sentences from the four systems was reported to be less than two hours.

Wagner recommends this method for MT evaluation by the user as it is less subjective than measuring post-editing time or rating intelligibility or accuracy using scoring scales and requires, therefore, only one person to assign error categories. The author does, however, note that any possible subjectivity can be decreased by comparing annotations across translations and by consulting other bilinguals to check agreement on error assignment.

3.4.8.3 Critical analysis

Using error classification for MT evaluation is the most objective method investigated so far in this chapter and such evaluations can, therefore, be carried out by only one annotator, although he/she must be bilingual. The method is informative for both users and developers and it is suitable for use with any text type. Moreover, the analysis of a small number of sentences is enough to compare different systems; Wagner's annotation of 27 sentences provided sufficiently discriminatory results.

Some element of subjectivity still exists, however, since opinions about which lexical items are appropriate or not in a given context can vary. Furthermore, some errors can be corrected in different ways, so more than one equally acceptable error assignment may be possible. For these reasons, in contrast to any of the other evaluation methods described in this chapter, it is preferable to use the same person to annotate all

translations of any given source text in order to maintain consistency. It is also advisable to obtain annotations from more than one evaluator, if only with a small sample, to test the validity of the scheme, the scope of interpretation and the level of inter-annotator agreement.

As Flanagan points out, a slightly different set of error categories is required for each language pair, to represent errors that typically occur. A good deal of initial preparation is, therefore, required to develop a classification if a suitable scheme does not already exist. This involves the detailed analysis of a sample of MT output, along with the source text, to record and classify all errors observed. Of course, once a scheme is developed, it can be reused time and time again.

Once a classification is ready for use, the annotation of errors (whether or not error correction is part of the process) is considerably more time-consuming than, for instance, using a scale to rate readability or fidelity; Wagner took almost two hours to annotate errors in 108 sentences, requiring over one minute per sentence. In contrast, in an evaluation of fidelity using an *n*-point scale (where the source and a human reference translation must be analysed) an intuitive judgement of a sentence can be made in a matter of seconds.

Weighting error categories is complex and little work has been done in this area. Flanagan divides errors into groups (Classes One, Two or Three) depending on how much each error type affects improvability or intelligibility. These groups can form the basis for weighting. Wagner simply counts the errors, but this means that instances of incorrect punctuation carry the same weight as, for example, wrongly translated words which have a far greater impact on quality. It is clear, therefore, that more work on error weighting for different evaluation purposes is required.

3.4.9 Evaluation using test suites

A test suite is "a collection of (usually) artificially constructed inputs, where each input is designed to probe a system's treatment of a specific phenomenon or set of phenomena." (Balkan, 1994). Within the TSNLP (Test Suites for Natural Language Processing) project, test suites are classified as monolingual, bilingual or multilingual; they can be based on linguistic phenomena found in a corpus or be artificially

constructed; test inputs can be words, phrases or sentences; they can be general or specific; domain independent or domain specific (Estival et al, 1994).

Test suites for MT tend to be used mainly by developers for glass-box, diagnostic and progress evaluations. As such, they are normally designed to test syntactic rather than lexical coverage; the latter being easily updateable. Their key properties, according to (Lehmann, Oepen et al., 1996) are (i) systematicity, (ii) control over data, (iii) inclusion of negative data, and (iv) exhaustivity. Test inputs are paired with their acceptable (human) translations and each one is used to test an MT system's correct treatment of a particular linguistic phenomenon. Consequently, each input must contain only one item of interest and other elements must not pose any translation problems. This helps the developer to pinpoint the cause when a translation error occurs. The fact that only one item is tested in each input means that its translation, in the majority of cases, is either acceptable or unacceptable, making this the most objective kind of evaluation.

Test suites can also be constructed for user evaluations, in which case, (Lehrberger and Bourbeau, 1988) suggest that test inputs should be taken from corpora that represent the intended use of MT. (King and Falkedal, 1990) also maintain that the frequency of each test item in its corpus of origin should be recorded and taken into consideration when conducting such an evaluation.

Test suites are often constructed to evaluate one particular MT system, so they are rarely publicly available and may not be suitable for other evaluations due to the specific phenomena they contain. Two particular projects addressed this lack of publicly available resources and developed shareable test suites, methodologies and tools for the evaluation of MT systems and other NLP applications. These are described below. Articles on the use of test suites for MT evaluation include (Heid and Hildenbrand, 1991), (Nerbonne et al., 1993), (Nyberg et al., 1994), (Dauphin and Lux, 1996) and (Koh et al., 2001). The latter is based on the JEIDA methodology described below.

3.4.9.1 JEIDA's test sets for quality evaluation of MT systems

This project was carried out by the Special Interest Group on Machine Translation, established by the Japan Electronic Industry Development Association. The work, described in (Isahara et al., 1994) and (Isahara, 1995), focuses on the evaluation of systems from a developer's point of view using test sets. The method claims to be

objective in terms of the evaluation process and the judgement of results. The English-Japanese/Japanese-English test suite comprises example sentences, their model translations and questions requiring yes/no answers, according to the behaviour of a chosen linguistic feature. MT and NLP system developers referred to grammar books to collect sentences covering all basic linguistic phenomena, along with a selection of items that are difficult to handle by MT systems. Two groups of test points were compiled, for (1) structural analysis (parts-of-speech, partial structure of sentence and sentence structure) and (2) structural selection (structural disambiguation and semantic disambiguation), each with a number of sub-categories. Inputs contained no unnecessary complexity.

An experimental evaluation of five English-Japanese MT systems was conducted with the aim of refining the initial set of test points and questions. The evaluation involved generating MT output of 309 SL examples, answering the questions, then counting the number of yes and no answers. Weighting according to frequency and significance was not taken into consideration. The number of correct answers given ranged between 53% and 80% and, although results were not seen as meaningful due to the lack of weighting, it indicated areas in which the five systems performed quite differently.

The final versions of the test sets were made publicly available for developers, researchers and users of MT. The English-Japanese test set comprises 770 items, and the Japanese-English version contains 730 test points. Grammatical explanations are also included for each linguistic feature.

3.4.9.2 TSNLP: Test Suites for Natural Language Processing

The aim of the TSNLP project was to survey existing test suites and their reusability and, based on findings, to compile a shareable database of test items, annotated with linguistic and other information, from which subsets of data could be extracted for the evaluation of various NLP applications. Tools were also developed for the construction and manipulation of test data for specific evaluation needs (Estival et al., 1994).

The TSNLP survey included an investigation of nine test suites, among which four had been designed for MT evaluation. These are described in Table 3.3, providing examples of the sizes and purposes of these resources.

Test suite	Purpose	Number of inputs	Origin of inputs	Approx. size of vocabulary
English Eurotra	For developers to test syntactic phenomena	955 English sentences	Corpus and artificially constructed	400 words
Aerospatiale	For users to evaluate commercial MT systems	1440 English-French items 346 French-English items	Corpus and artificially constructed	800 words
Systran	For users to test verb codings	853 French-German items	Artificially constructed	800 words
IAI	For user evaluation of LOGOS	2481 German-French/ German-English items	Artificially constructed	Very large

Table 3.3: TSNLP Project: summary of test suites for MT

Following this survey, the subsequent DiET project (Diagnostic and Evaluation Tools for Natural Language Applications) compiled a database of comparable test items in English, French and German. For each language, around 5,000 test inputs were systematically constructed and categorised under a large number of headings (eg. negation, structure of NPs, auxiliary verbs) and sub-headings. The items are designed to test mainly syntactic, but sometimes morphological and extra-grammatical coverage (eg. punctuation and the handling of numerical data), and the treatment of ill-formed input. Ambiguous words are avoided unless the item of ambiguity is under examination. Users can add their own test data, such as translations of test inputs for MT evaluation, without affecting the underlying model. Information on the tools can be found in (Klein et al., 1998).

3.4.9.3 Critical analysis

The use of test suites is the most informative evaluation method for developers, enabling the diagnosis of specific syntactic errors for remedial action. Furthermore, the projects described above show that test suites can also be useful for the evaluation of commercial systems from a user's perspective. However, unlike all other human evaluation methods explored in this chapter, test suites are not designed to evaluate lexical coverage, but only the correctness of grammatical structures. For this reason, their use is inappropriate for many user needs.

A further drawback concerns the time and complexity involved in the construction of a test suite; expert linguistic knowledge is required, with constant reference to grammar

books or corpora, for the systematic construction of test inputs. Hundreds or thousands of test points are needed, as shown in the examples above, along with a translation of every item. Furthermore, a different set of test points is required for each language pair and direction and for each text type and domain.

Once a test suite is compiled, however, it becomes a valuable resource that can be reused and adapted as needed. A major advantage is that evaluations can be carried out by one person, as the method is objective, and little time is required to make a decision on the acceptability of each specific test point. Bilingual knowledge is, however, required of the developer/evaluator although, arguably, a monolingual native speaker of the target language could answer the yes/no questions used in the JEIDA methodology.

Further research is required into the applicability of test suites for black box MT evaluation from a user's perspective, and the weighting of test points needs further investigation. The objective nature of test suite evaluation means that translated test points can be assessed automatically, using an algorithm to automatically compare the machine translated item (or required fragment thereof) with the expected output. This would enable developers to quickly identify the incorrect items, and users to obtain an automated score for the test set in question. This premise will be investigated for the design of an automated error detection system for the evaluation of MT systems for the current research.

3.4.10 Human evaluation methods: conclusions and directions for this research

The above investigation of human evaluation methods now enables the informed selection of a method (or methods) to use when evaluating translations for this research. It also provides information that may be useful for the development of an automated approach. The aim of this work, as stated in Chapter 1, is to develop a new automated MT evaluation method and, in order to validate or reject any new approach, human judgements of MT output must be obtained, against which automated scores can be compared. The granularity and nature of these human judgements and the practicalities involved in the design and implementation of the evaluation must be carefully considered in view of financial and other constraints.

As the intention is to develop an automated method that requires access only to the MT output, our priority will be to test whether automated scores can predict human judgements obtained in the same way, without access to the original or to any human translation. In addition, it will be interesting to investigate whether such automated scores can also predict human judgements of fidelity, obtained with reference to the ST or an expert human translation. A method from each of these two categories must, therefore, be chosen.

The above investigation of human evaluation approaches has highlighted many factors that must be taken into account when selecting a method appropriate for particular needs. In the light of these findings, eight priorities have been identified:

1. Scores must be reliable, even if more than one judgement is required, ie. the fewer variables affecting inter-evaluator agreement, the better;
2. Judgements should be provided at sentence (or segment) level rather than text level, offering the possibility for fine-grained comparisons between human and automated scores, should they be required;
3. It should be possible to conduct the evaluations with monolingual subjects, as they are more readily available and less expensive;
4. No expert linguistic knowledge or experience/skills in a particular domain should be needed. This would make the recruitment of evaluators more difficult;
5. Methods should be usable with most text types and domains, enabling the evaluation of a variety of genres which reflect real use of MT systems;
6. Methods should provide fine-grained judgements, even with output of very poor quality, enabling the comparison of systems of any level of performance;
7. Methods should be applicable to the evaluation of human translations, enabling us to quantify differences between MT and humans and to make comparisons between MT systems more meaningful;
8. Subjects should be able to conduct evaluations at a computer, eliminating printing and paper costs and, importantly, enabling the automated computation of results.

The nine evaluation methods investigated in this chapter are listed in two categories in Tables 3.4 and 3.5. The priorities fulfilled by each method are then indicated, based on information and opinions derived from our study.

Priority	Readability using n-point scales	Readability by cloze test	Readability by reading time	Performance -based evaluation
Method must provide reliable scores	●	●		
Texts judged at sentence/segment level	●			●
Monolingual subjects can be used	●	●	●	●
No expert linguistic knowledge/skills required	●	●	●	
Method usable with most text types/domains	●		●	
Suitable for any quality of MT	●		●	
Method can be applied to HTs for comparison	●	●	●	●
Evaluation can be conducted at a computer	●	●	●	

Table 3.4: Priorities for evaluation methods without access to ST or HT

Preference	Fidelity using n-point scales	Informative -ness by questionnaire	Post-editing effort	Error analysis	Test suites
Method must provide reliable scores	●			●	●
Texts judged at sentence/segment level	●				●
Monolingual subjects can be used	●	●			
No expert linguistic knowledge/skills required	●	●			
Method usable with most text types/domains	●	●	●	●	●
Suitable for any quality of MT	●	●	●	●	●
Method can be applied to HTs for comparison	●	●	●	●	●
Evaluation can be conducted at a computer	●	●	●	●	●

Table 3.5: Priorities for evaluation methods with access to ST or HT

Findings presented in Tables 3.4 and 3.5 show that the evaluations of readability and fidelity using n-point scales are most appropriate for this research. These methods may be more subjective than error analysis or the use of test suites, requiring more than one judgement per segment, but they require no bilingual or linguistic knowledge on the part of the evaluators. Furthermore, these methods can produce fine-grained intuitive judgements which a new automated evaluation method will attempt to reflect. The next

step in terms of human evaluations involves the selection of texts, decisions on the number of judgements per segment, the type of evaluators to be used, and the development of scoring metrics and evaluator packs. These issues are discussed in Chapter 4.

3.5 Automated evaluation methods

The focus of the remainder of this chapter is on automated MT evaluation methods, which have emerged more recently due to the time and expense required for human evaluations. Such a critical investigation will enable us to establish work already carried out in this relatively new area so that (a) the strengths and weaknesses of existing approaches can be analysed to provide pointers for this research and (b) approaches that remain uninvestigated can be considered as a basis for the development of a potential new method.

A great deal of research and testing is required for the development of automated evaluation metrics, as methods are only validated when strong correlations are found between automated scores and intuitive human judgements. However, once a method is developed, it can be used repeatedly to produce objective scores quickly and cheaply, and evaluating a large sample of translations requires little more time than the processing of a small number of texts. Moreover, due to the objectivity of automated approaches, the training effects involved in manual evaluations are eliminated, so that a single program can compute scores for many translations of the same source sentence. This ability to produce instant results is ideal for developers who need to perform frequent evaluations to monitor system improvements, and is also desirable for users who need to quickly and cheaply compare different outputs.

Approaches to automated evaluation can be divided into six categories, summarised in Figure 3.11. As with the human evaluations shown in Figure 3.1, evaluations are divided into those which do and do not require access to the source text or human translations.

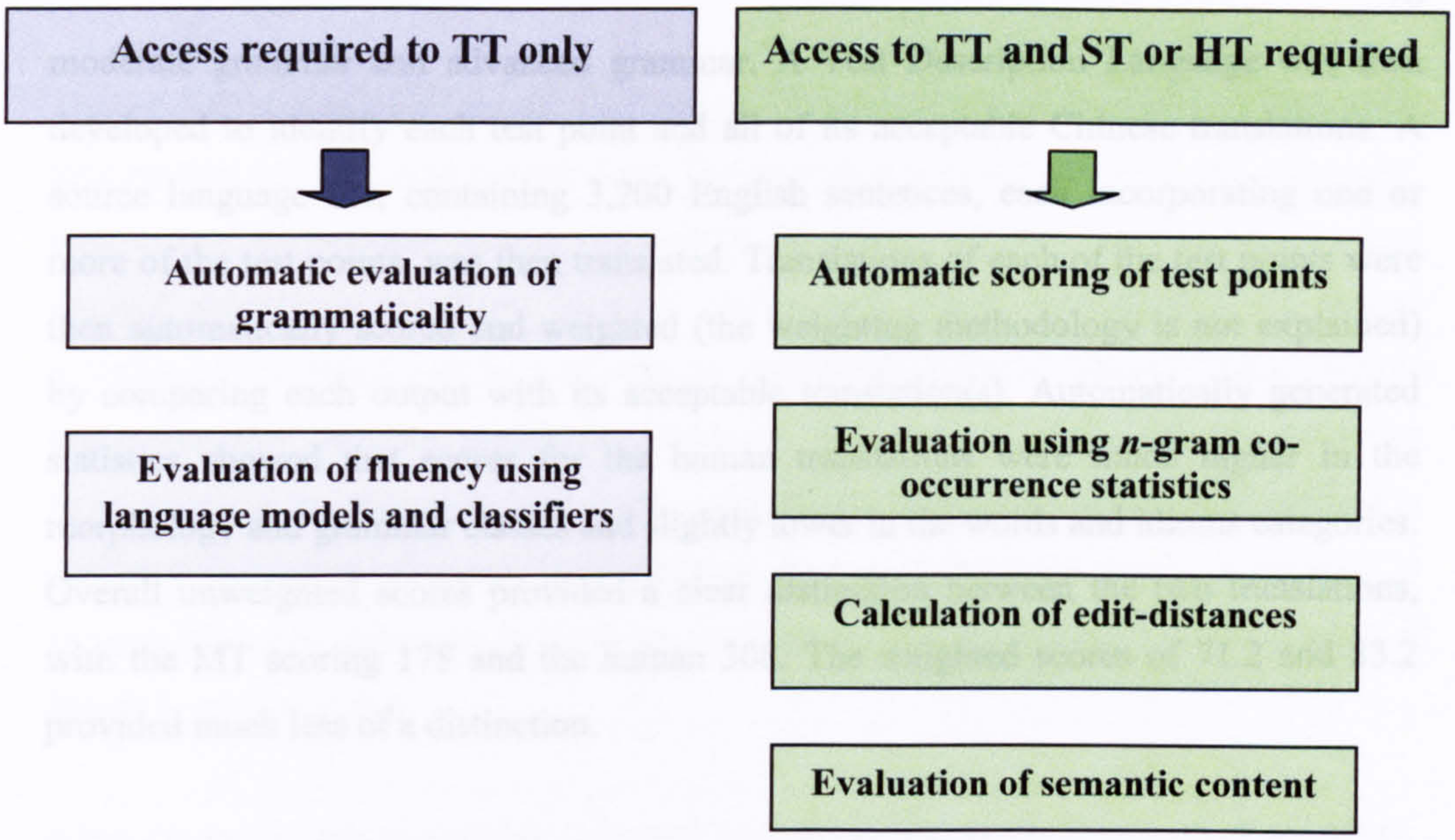


Figure 3.11: Types of automated evaluations

In 3.5.1 we describe one of the earliest attempts of automated evaluation, based on the automatic scoring of test points. Section 3.5.2 focuses on the use of n-gram co-occurrence statistics and 3.5.3 investigates the calculation of edit distances between human and machine translations to rank MT systems in accordance with human judgements. In 3.5.4 we explore the use of NLP tools to evaluate grammaticality and semantic content and in 3.5.5 we analyse the use of language models and classifiers which are trained to identify the most fluent translation from a set of outputs. Finally, section 3.5.6 focuses on investigations of potential automated methods, which have been explored manually, based on metrics for the objective scoring of particular translated features.

3.5.1 Automatic scoring of test points

This early attempt to automate the evaluation process (Shiwen, 1993) was developed at Peking University to evaluate English-Chinese translations. The method was tested with sentences translated by the TRANSTAR MT system and by a Chinese native speaker with a very basic knowledge of English.

A file of several hundred short test inputs was created, each one involving a specific problem for English-Chinese translation. The test points were divided into six weighted classes for subsequent scoring: words, idioms, morphology, elementary grammar,

moderate grammar and advanced grammar. A Test Description Language was then developed to identify each test point and all of its acceptable Chinese translations. A source language file, containing 3,200 English sentences, each incorporating one or more of the test points, was then translated. Translations of each of the test points were then automatically scored and weighted (the weighting methodology is not explained) by comparing each output with its acceptable translation(s). Automatically generated statistics showed that scores for the human translations were much higher in the morphology and grammar classes and slightly lower in the words and idioms categories. Overall unweighted scores provided a clear distinction between the two translations, with the MT scoring 178 and the human 308. The weighted scores of 71.2 and 83.2 provided much less of a distinction.

3.5.1.1 Critical analysis

This method is informative for developers, as results highlight areas for system improvement. However, it could also be used for comparison evaluations from a user's perspective if automated scores were found to be predictive of human judgements. (We assume that this would be more likely with a very large number of weighted test points.) The method can be used to evaluate human as well as machine translations very quickly, and is appropriate for different text types and domains.

The main drawback is that a new set of test points (including all possible translations) must be written for each language pair and, as described in 3.4.9, this process is very complex and time-consuming. Furthermore, with some TLs, such as Chinese, numerous translations of the same sentence are acceptable and must be included. However, once a test suite is of sufficient size, it can be used repeatedly and modified or extended as required.

3.5.2 Evaluation using n-gram co-occurrence statistics

Methods such as BLEU (Papineni et al., 2001) and its adaptations (mentioned below) are based on the assumption that the closer a machine translated sentence is to an expert human translation, the better it is. Sentences are automatically evaluated by measuring their proximity to one or more human reference translations. Scores produced by these methods have proven to correlate highly with human judgements of adequacy and fluency (Papineni et al., 2001, 2002), (Doddington, 2002), (Coughlin, 2003), (Leusch et al., 2003), (Lavie, 2004), (Lin and Och, 2004).

The BLEU method (BiLingual Evaluation Understudy) was designed to be quick, inexpensive and language independent. The approach is based on the "word error rate" metric used to evaluate speech recognition programs, but was adapted to take into account acceptable variations in word choice and syntax. Translations are scored according to modified n-gram precision, by comparing unigram, bigram, trigram and 4-gram matches between candidate sentences and one or more human reference translations (in an attempt to account for variation by human translators). The method is designed to capture adequacy, by measuring the use of the same words (unigrams) in the candidate and reference translations, and fluency, by accounting for the number of longer *n*-gram matches.

After automatic pre-processing (removal of case information, treatment of numbers, etc.,) each machine translated sentence is compared to the available reference translations. Words that do not appear in any of the reference translations or that occur more frequently in the candidate translation than the reference are penalised. Penalties are also applied to sentences that are significantly longer or shorter than their reference translations.

In an initial evaluation of the method (Papineni et al., 2001), involving Chinese-English translations of fifty sentences by three MT systems and two humans, automated scores (based on two reference translations) and human scores were compared. A high correlation (up to 0.99) was observed between the BLEU and human scores; the method was found to accurately estimate large differences between machine and human translations and to be sensitive to small differences in performance between systems.

Further evaluations of the method are reported in (Papineni et al., 2002) for additional language pairs: Arabic-English (19 documents), French-English (all 100 DARPA texts) and Spanish-English (100 DARPA texts). Correlation coefficients between the automated scores and human judgements across systems varied between 0.94 and 0.9958.

The BLEU method was adapted for the NIST MT evaluation campaign (Doddington, 2002) and is currently used for TIDES sponsored MT research (NIST, 2005). The NIST method varies from BLEU in terms of text pre-processing (eg. case is not normalised, so

incorrect case affects scores), n-grams are weighted based on their frequency and only trigrams are used. The BLEU and NIST approaches were compared using DARPA 1994 texts for all three language pairs (French, Spanish and Japanese into English) and a Chinese-English corpus of 80 documents. The NIST method correlated more highly than BLEU with human scores for adequacy for all four language pairs and for fluency with the Chinese-English pair (Doddington, 2002). However, much lower correlations were observed between automated and human scores for the human translations (70.5% for adequacy and 16.6% for fluency) indicating that small differences between professional translators are less well characterised by n-gram co-occurrence statistics.

(Babych, 2004) and (Babych and Hartley, 2004) extended the BLEU approach with weights based on the statistical significance of lexical items. The idea is based on Legitimate Translation Variation (LTV), which takes into account the fact that some pieces of information (eg. named entities) are more important than others (and should carry a higher weight) and that certain other information is subject to greater, legitimate variation. BLEU scores are adjusted according to weights which are automatically generated based on the frequencies of n-grams in the candidate text compared to their frequencies in a larger corpus. This method was compared with BLEU using the French-English DARPA 1994 texts. It was found, in particular, to measure adequacy more accurately than BLEU for output from a statistical MT system and the stability of scores was improved using only one reference translation.

Other adaptations of BLEU include NEVA: N-gram EVAluation (Forsbom, 2003) and ROUGE-L and ROUGE-S (Lin and Och, 2004), based on the longest common subsequence statistics between candidate and reference translations, and on skip-bigram co-occurrence statistics (pairs of words in their sentence order, allowing for gaps). (Turian et al., 2003) developed the General Text Matcher (GTM), an online tool which produces scores based on recall, precision and the F-measure; the latter was found to be more accurate than either BLEU or NIST. Some research has explored alternative methods for the computation of scores using BLEU, such as (Lavie, 2004) who places more weight on recall than precision to obtain better correlations with human judgements.

3.5.2.1 Critical analysis

The measurement of proximity to human reference translations has proven to be a quick and reliable evaluation method, once a reference corpus is created. The method is particularly informative for users, as scores from a set of fifty sentences have been shown to correlate highly with human judgements. Furthermore, although only raw scores are generated, rather than information on particular errors, it can help developers to pinpoint particularly badly translated sentences.

This method of evaluation does, however, have several drawbacks. First, it relies on a corpus of human reference translations, comprising at least one, and preferably several, translations of each source sentence. These need to be produced for every language pair to be evaluated and are often reused in subsequent evaluations to save time and expense, even perhaps in cases where different kinds of texts would be more appropriate for the needs of the stakeholders in question. Second, the majority of adaptations of this approach are based on the assumption that all words are of equal importance. So, for instance, a superfluous determiner (with little or no effect on fidelity) would be penalised to the same extent as an incorrectly negated word, which completely changes the meaning of a sentence. Furthermore, synonyms are also penalised, even in cases where they are acceptable, if they are not found in the reference translation(s) used.

It has been shown that these methods are sensitive to small differences in MT quality, and that scores accurately predict human quality judgements. However, correlations between human and automated scores are much less reliable for human translated texts, or for evaluating machine or human translations at the sentence level. (Gamon et al., 2005), for instance, found a correlation of just below 0.42 between BLEU and human scores, based on 500 sentences translated from computer manuals and online help files.

Another major drawback concerns the comparison of results from different evaluations. For example, the greater the number of reference translations used, the higher the automated score will be, so comparing results from evaluations using different numbers of references is meaningless. Furthermore, scores are not comparable across different TLs and should be calibrated separately across languages using human evaluation scores (Babych et al., 2005). In this work it was observed that the number of n-gram matches depends on whether the TL allows for a greater or smaller degree of variation for different forms of the same word and for word order. It was noted that this can be a

major issue for heavily inflected Romance languages, such as French, Italian and Portuguese, or languages with a 'free' word order, such as German.

(Culy and Riehemann, 2003) questioned the reliability of scores for different text types, observing that disfluent machine translations can receive better scores than completely fluent human translations in an evaluation of translated extracts from the bible and a work of fiction: very literal human reference translations gave lower scores to the more 'free' human translations and higher scores to the more literal MT. Furthermore, their findings illustrate how scores can vary immensely using different sets of reference translations. Their findings also showed that a low n-gram score does not necessarily indicate a bad translation, but that a high score is likely to be indicative of a good translation.

3.5.3 Ranking MT by computing edit distances

Using edit distances between machine translated sentences and human reference translations has formed the basis of a number of MT evaluations in recent years. Automatic calculations are based on the Levenshtein distance (Levenshtein, 1965) and the formula can be adapted as required. This is based on the number of insertions, deletions and substitutions required to convert one string into another; in this case, a machine translated sentence into a human reference translation. The greater the similarity to a human translation, the better the MT is assumed to be, and the smaller the edit distance. Editing units can be phrases (Thompson, 1991), characters (Su, 1992), words (Takezawa et al., 1999), (Sugaya et al., 1999), (Yasuda et al., 2000, 2001), (Forsbom, 2003), morphemes and words (Akiba et al., 2001) and blocks of words (Leusch et al., 2003).

(Niessen et al., 2000) developed a semi-automated evaluation tool called EvalTrans, incorporating an automated string edit distance metric. The edit distance or Word Error Rate (the length-normalised Levenshtein distance) between a machine translated sentence and one or more human reference translations is automatically measured. A 'Subjective Sentence Error Rate' (SSER) is then assigned by a human using a scale of 0-10, with access to the ST. The ST sentence is then stored in a database with the human reference translation, the machine translation, the human score and the Word Error Rate (WER). All associated data can be viewed and manipulated via a graphical user interface.

An SSER for a new machine translated sentence can then be estimated automatically by comparing its WER to those of previous translations of the same sentence in the database and by computing the average SSER of the closest translations to the candidate. The automatic score can then be considered and amended if necessary by a human, and new information fed back into the system. Tests showed that once sufficient data is stored, differences between estimated scores and human SSERs are minimal, but success is dependent on the number of translations already contained in the database.

A similar tool was developed by (Tomás et al., 2003) to evaluate Spanish-Catalan MT. This freely available online resource enables the user to compute various different measures of edit distance, including word error rate (WER) with one reference translation, multiple reference WER, sentence error rate (SER) and subjective sentence error rate (SSER).

A fully-automatic evaluation method based on multiple edit distances was devised and tested by (Akiba et al., 2001) to help MT developers who needed to conduct frequent evaluations of the same system. The RED approach (Ranker based on Edit Distances) automatically ranks machine translations of the same ST sentence based on a set of sixteen different edit distances to multiple reference translations. The different edit distances are based on calculations dependent on, for example, all words or only content words, defined keywords, correct parts-of-speech, actual words or semantic codes (to avoid penalties for acceptable synonyms). A set of edit operations was first defined, and a cost assigned to each one. Each edit distance is calculated as the sum of the costs in the cheapest chain of operations when transforming a sentence into the most similar reference translation.

In the learning phase, edit distances are computed for MT sentences in a training set, and each is then assigned a rank in relation to all other translations of the same source sentence by at least three human evaluators. This information is then encoded into a sixteen-dimensional vector from which a classifier of ranks is learned in the form of a decision tree. In the evaluation phase, ranks are computed for new sets of MT sentences using the learned decision tree.

The use of multiple edit distances was found to be more accurate than methods based on single edit distances. However, a test involving 343 sentences, each with 26 reference

translations, showed that either 13 or 19 human translations of each sentence provided the closest match to ranks assigned by humans.

3.5.3.1 Critical analysis

The computation of edit distances between machine and human reference translations has been shown to rank MT output in accordance with human judgements. However, these methods have many disadvantages. A major problem is the requirement of a large number of human reference translations to provide a reliable ranking of systems. In addition, (Akiba et al., 2001) require a tagger to annotate parts-of-speech, a database of semantic codes, a DT learner and a large set of training data to produce the most accurate results. A large number of translated sentences must also be evaluated by humans, so that mappings between edit distances and human ranks can be learned.

In terms of the reliability of ranks, these methods are very sensitive to word order (Forsbom, 2003), causing good quality translations to obtain lower ranks when the word order differs considerably between MT and reference translations. This problem was partly solved by (Leusch et al.,;2003) who modified the algorithms to move blocks of correctly ordered text at the same cost as single words. However, evaluations in which the TL word order is more flexible (eg. Chinese) means that results will not be comparable with other languages, unless a very large number of reference translations is provided.

Furthermore, like methods based on n-gram co-occurrence, the measurement of edit distance considers words, or word groups, to be of equal importance, so substituting an incorrect content word can carry the same penalty as, for example, replacing an acceptable synonym.

3.5.4 Using NLP tools to evaluate grammaticality and semantic content

A limited number of investigations have focused on the feasibility of computing scores or ranks for MT based on the performance of Natural Language Processing tools when handling translated sentences. Some work has been influenced by (Jones and Rusk, 2000) who extracted numeric values from various pieces of software, such as parsers and NLP tools, to measure mutual information to reflect the degree to which they could successfully handle machine translated sentences. Their objective was to define a

scoring function for the quality of English to attempt to machine-learn a good translation grammar for system development.

(Rajman and Hartley, 2001, 2002) explored three methods, using parsers and a vectorial model for semantics, in an attempt to automatically predict ranks assigned by humans to machine and human translations of the same texts. Data from the DARPA 1994 corpus was used to test their potential scoring methods:

The C-score was devised to measure the grammaticality of the translations using a stochastic context-free parser to produce a syntactic bracketing for each sentence in a text. Any words not found in the lexicon were associated with all possible open parts-of-speech. The average bracketing coverage was then computed to give the mean number of words in a maximal bracketing for each sentence. This number would represent the grammaticality of the sentence. Values were then normalised according to sentence length and the C-score was calculated by finding the mean sentence value for each text.

The X-score was also devised to measure grammaticality. The Xerox shallow parser XELDA was used to produce the syntactic dependencies for each sentence. In total, 22 different dependencies were produced. A dependency profile was then computed for each text, comprising the number of occurrences of each of the 22 dependencies. A formula was then selected from several possible candidates to produce a single score: X-score = (#RELSUBJ+#RELSUBJPASS-#PADJ-#ADVADJ).

The D-score was developed to measure the preservation of semantic content in a translation. A vectorial model for semantics and a large corpus of aligned translations were used. The position of the source text in the semantic vector space defined by a source language reference corpus was compared with the position of the target text in the semantic vector space defined by a target language reference corpus. If the semantic content was well preserved in a translation, the similarity matrices for the source text in the source vector space and the target text in the target vector space should be similar. The distance between the two matrices was used to produce a score to indicate how much of the semantic content had been preserved.

No strong correlation was found between any of the three automated ranks or between any set of human DARPA scores (for fluency, adequacy and informativeness) and automated ranks for all systems. Furthermore, a large difference was observed between the human and automatic rankings of the human translations. It was, therefore, decided to explore correlations between the overall rankings of the MT systems alone. After extensive calculations, it was concluded that the X-score was the best predictor for fluency, but that this method should be tested on larger corpora. The D-score was the best predictor for adequacy and informativeness.

Further work was carried out to test the robustness of the X and D-scores (Rajman & Hartley, 2002), the C-score being insufficiently predictive. Human translations were excluded from the experiments. Five machine translations of twenty French source texts from the DARPA corpus, and translations from one additional system, were ranked by human subjects. While the D-score showed a relatively weak correlation, the X-score proved to be a very good predictor of the human rankings.

3.5.4.1 Critical analysis

Of the above three methods, only the X-score is discussed here as it was shown to be the only good predictor of human rankings. Computing the X-score has the major advantage of not requiring human reference translations. Furthermore, it is usable with any language pair in which the TL is English, and the formula could be adapted to evaluate other target languages.

However, there are several drawbacks: firstly, significant differences between human and automated scores for the human translations indicate that the method is suitable only for the evaluation of machine translations, preventing comparisons between the quality of MT and human output. Secondly, the method places systems in rank order, but does not produce finely differentiated scores. This prevents any fine-grained investigation of the extent to which output from different systems varies in quality. Thirdly, the X-score rates the grammaticality of sentences, but does not account for the preservation of the content of the original in the translations. Finally, although the approach is suitable for evaluations from a user's perspective, enabling the prediction of intuitive human judgements, it is not so useful for developers, as reliable rankings are achieved at system level with several texts, rather than at sentence level.

3.5.5 Using language models to rate fluency

A small number of research projects have focused on the use of language models, not specifically to score or rank MT systems, but to identify the most fluent sentence from a number of translations of the same ST. This approach could be adapted to produce scores or ranks for alternative outputs. Methods are based on the assumption that the most fluent sentence is the best translation and no access is required to the ST. Research in this area includes (Kaki et al., 1999), (Callison-Burch and Flournoy, 2001), (Corston-Oliver et al., 2001), (Akiba, 2002), (Kulesza and Shieber, 2004) and (Gamon et al., 2005).

(Callison-Burch and Flournoy, 2001) developed a program that automatically selects the best (most fluent) translated sentence from a set of candidates using a trigram language model. This was developed to enable a translation provider to offer the best machine translations to its clients, comprising a combination of the best sentences from different systems rather than the whole output from one system. The method is based on the assumption that if an MT system produces a fluent translation, it is likely that the engine successfully analysed the ST, increasing the likelihood of the correct transference of meaning.

A statistical language model for English was built, based on a corpus of just over two million words and a simple trigram model was generated. To assign a probability that a machine translated sentence is English, the program computes the number of occurrences of each word, bigram and trigram from that sentence in the whole corpus. The probability of each trigram is then smoothed with the probabilities of the bi-grams and unigrams to compute a probability for the sentence.

The method was evaluated using output from four Japanese-English and two French-English MT systems. An English native speaker rated the fluency of each sentence on a scale of 1-4. The program was awarded a point for each time the automatically rated best sentence corresponded to the highest ranked sentence by the human. The tool chose the best translated sentences 74% of the time for Japanese-English and 84-94% of the time for French English depending on text type. In a further test, applied to French as the TL, and using a language model based on a 1.1m word corpus, the tool selected the most fluent sentences 67% of the time.

(Corston-Oliver et al., 2001) at Microsoft Research developed a machine learning approach to evaluate the wellformedness of MT output, building classifiers that learn to distinguish human translations from MT. This method was designed for developers to track system improvements over time. Again, the method is fully automated and, based on the same assumptions as (Callison-Burch and Flournoy, 2001), it is SL independent and takes no account of the transfer of content.

Language models were built from 200,000 English sentences taken from computer software manuals and online help files, representing the kinds of texts that the MT system was required to translate. Classifiers were then trained using 90,000 Spanish-English machine translations and 90,000 human reference translations. Linguistic features were selected for a classification task, based on an analysis of failures in system output. A syntactic parse was performed for each sentence to automatically extract 46 features, including perplexity measures, branching properties of the parse, density of function words and other parts-of-speech and constituent lengths (noun phrases, prepositional phrases etc.). Decision trees were then constructed based on the selected features.

The method was tested on 20,000 sentences of Spanish-English MT and human translations, achieving an accuracy of 82.89%. The authors suggest that the approach can be used to evaluate MT systems by evaluating the accuracy of the classifier with new data: the better the MT output is, the more human-like it will be, and the accuracy achieved by the classifier will reduce.

The above approach is currently used at Microsoft Research to guide MT system development, and further experiments were reported by (Gamon et al., 2005). The latter emphasised the need to evaluate large amounts of text quickly and to identify particularly bad sentences for rapid post-editing. The use of language models trained on a domain-specific corpus in the TL enables the quick identification of poorly translated sentences in the absence of human reference translations. Language models were trained with 1.5 million French sentences and support vector machine (SVM) classifiers were trained with 460,000 human and machine translated English-French sentences, all in the software domain. The approach was tested on 500 held-out sentences; language model perplexity scores and SVM class probability scores were combined to produce the maximum correlation with human judgements of fluency. The combined method

achieved a correlation of 0.42 at sentence level, which outperformed BLEU very slightly. In spite of this low correlation, the authors find the method to be a good indicator of particularly badly translated sentences.

3.5.5.1 Critical analysis

Based on the assumption that the most fluent translation is the best one, like (Rajman and Hartley, 2001, 2002) these approaches have the advantage of not requiring human reference translations, and being easily adaptable for any target language. However, both methods require the initial processing of large corpora (of comparable texts or human translations) before evaluations can be conducted. In spite of these initial set up costs, once the preparation is complete, automated evaluations can be performed very quickly provided that candidate texts are in the same domain and target language as the corpus.

The accuracy of these methods is relatively high at the sentence level, although (Callison-Burch and Flournoy, 2001) report correspondences only between human and automated ranks for the most fluent sentence from each set of translations. Furthermore, evaluations have not incorporated human candidate translations, so it is currently unclear as to whether these methods could provide accurate comparisons between human translations and MT output.

While the first method (Callison-Burch and Flournoy, 2001) was designed to be informative for users, to identify the best translated sentence from a selection, the approaches tested by (Corston-Oliver et al., 2001) and (Gamon et al., 2005), were devised for developers and post-editors, enabling the quick identification of particularly bad sentences. In spite of their original purposes, both approaches could be adapted to produce sentence-level scores, which could be informative for users and developers, provided that there is a strong correlation with human judgements.

3.5.6 Other experiments with potential automated methods

Several published articles have focused on the manual analysis of different automatically measurable features in MT output, to determine whether or not automatic scores based on these features might correlate with intuitive human quality judgements.

(White and Forner, 2001), for example, used texts from the DARPA corpus to examine the behaviour of noun compounds in French-English and Spanish-English translations. Each translation was manually scored (with reference to the source text) according to the number of correctly translated noun compounds over the total number contained in the text. To maintain consistency across evaluators, noun compounds were scored in terms of syntactic correctness, regardless of lexical choice or the insertion or omission of determiners. Despite some deviations, the number of correct noun compounds was found to increase with the DARPA scores for adequacy for both language pairs.

In a similar experiment, (Reeder et al., 2001) explored the feasibility of predicting human scores based on the number of correctly translated named entities in the MT output. A sample of Spanish-English translations from the DARPA corpus was used to manually score the correct translation of named entities, including proper names, geographical features, dates and monetary amounts. These scores were compared with the existing DARPA fidelity and intelligibility scores to discover any possible correlation: over two thousand tagged named entities in the MT output were aligned with their corresponding expert human translations. Exact matches were given a score, along with slight mismatches which involved differences in capitalisation, the translation of a digit by a word (or vice versa) and the use of diacritics. The results were not as encouraging as hoped as only a weak correlation was found between mean scores for correct named entities and the DARPA scores. It was found that relaxing certain constraints aided some systems at the expense of others, indicating the need for further research in this area. It was concluded that additional scoring techniques should be considered, such as the analysis of relationships between named entities and technical and other specialised terminology.

Work reported in (Miller and Vanni, 2001) and (Vanni and Miller, 2001, 2002) involved the selection of qualities from the ISLE framework, for which evaluation metrics were devised and tested with a view to automating some or all of the methods. The priority here was to develop approaches, whereby scores could be mapped to particular information processing tasks, predicting which tasks would be performable with the output. This follows work by Church and Hovy (1993) and White and Taylor (1998) on task-based evaluation, which draws on findings that certain information handling tasks using MT output are more tolerant of errors. A second aim was to provide diagnostics to

help researchers and developers improve systems, rather than to test correlations between automatable metrics and human scores.

Two evaluators used seven different metrics to evaluate three translations of two Spanish texts from the DARPA corpus. The same metrics were then used to evaluate three translations of one Japanese text to test their validity with a language pair in which the source language is structurally very different. The chosen qualities and metrics are as follows:

- **Coherence** ("the degree to which the reader can define the role of each individual sentence (or group of sentences) with respect to the text as a whole" (FEMTI, 2005)): the evaluation method drew on a simplified version of Rhetorical Structure Theory (RST), described in (Mann and Thompson, 1988). Texts were scored according to the percentage of sentences to which some RST function could be assigned.
- **Clarity**: target sentences were rated on a 4-point scale. The text score was the mean sentence score.
- **Syntax**: this score was based on the minimum number of syntactic corrections (such as replacements, additions and deletions) required to render the translation grammatical. The text score was the ratio of the number of corrections per sentence to the number of words.
- **Morphology**: the morphology score was the total number of morphological corrections needed, divided by the total number of inflectable words in the text.
- **Dictionary update**: the percentage of untranslated words was calculated.
- **Terminology**: the terminology score was the percentage of domain-specific words correctly translated.
- **Names**: a score was given for the percentage of proper names correctly translated.

Relative rankings between the two evaluators were found to be consistent in all cases apart from the morphology evaluation. The dictionary update, terminology and name tests were found to be very clear-cut, objective measures and scores from both subjects were identical, indicating that these methods are suitable for automation.

3.5.6.1 Critical analysis

Work by (White and Forner, 2001) and (Reeder et al., 2001), shows that the identification of a single phenomenon, such as correctly translated noun compounds or named entities, is insufficient to produce scores that can reliably predict human judgements. Furthermore, (Babych and Hartley, 2004) showed that named entity recognition in MT does not produce scores that correlate strongly with either intuitive human judgements or BLEU scores; different MT systems handle certain linguistic features more successfully than others, so scores from an isolated feature are unlikely to represent the quality of the output as a whole. Furthermore, the number of noun compounds or named entities in a text depends, to a large degree, on the text type and subject domain, and some texts contain no named entities at all, making this kind of evaluation meaningless. We assume that more finely differentiated scores would be generated from texts containing a large number of the selected features.

Calculating the number of correct translations of named entities or noun compounds is objective, as these have fewer translation possibilities than many other linguistic features and, as such, their identification is easily automatable. However, the method may not be resource-light: it would require a large bilingual lexicon against which source and target text elements could be matched, and a different lexicon for each language pair (and potentially, for each subject domain) to make the method more tractable.

(Miller and Vanni, 2001) and (Vanni and Miller, 2001, 2002) suggest a number of possibilities for automated evaluation, but do not focus on how methods could be automated. However, we assume that the source text or a human reference translation would be required for most metrics, with the exclusion of the coherence and clarity evaluations which focus on the target text alone. The automated evaluation of syntax and clarity could be performed by adapting methods described in 3.5.3 and 3.5.5 respectively.

The scoring of (untranslated) terms for dictionary update, terminology and names is objective (although acceptable variations must be recognised) and is, therefore, automatable in much the same way as methods suggested above for the detection of noun compounds and named entities. Domain-specific lexicons can be compiled by extracting terminology from human reference translations, either manually or

automatically. However, correlations between human and automatable scores remain uninvestigated and it is clear that some features may appear more frequently in some texts than others, making the detection of some phenomena inappropriate for some text types.

The correct translation of named entities, noun compounds and terminology is important, as it makes MT output useful for tasks such as information extraction, which is an increasingly important use for MT. Furthermore, the flagging of untranslated terms for dictionary update is also useful for the evaluation of a system's lexicon. The automatic scoring of a greater number of linguistic phenomena, including those investigated here, would be useful to explore the possibility of stronger correlations with human scores using a combination of features. Focusing on a single linguistic phenomenon is of limited use to developers, but detecting the correct translations of many different phenomena may prove informative for both users and developers.

3.5.7 Automated MT evaluation methods: conclusions and directions for this research

An investigation of automated evaluation methods has enabled us to ascertain work already carried out in this field so that unexplored areas may be identified for further research. The approaches described above represent some of the most important work to date in this field, and many of these projects are continuing. However, each method has certain drawbacks: approaches based on the computation of n-gram co-occurrence statistics or edit distances rely on large numbers of human reference translations, which are expensive to produce; some methods were designed for particular purposes, and have not been shown to produce scores that correlate with human judgements of the same translations. These include experiments by (Shiwen, 1993) and (Callison-Burch and Flournoy, 2001). Some approaches cannot reliably predict human judgements of human translations (Rajman and Hartley, 2001) and others are not appropriate for the evaluation of certain kinds of texts (Reeder et al., 2001).

Many areas remain unexplored, and research into new approaches to automated MT evaluation and the adaptation of existing methods are essential so that further progress can be made. A clear understanding of research to date now enables us to identify directions for this research and facilitates the compilation of a list of desiderata for a new method.

3.5.7.1 Desiderata for automated MT evaluation

Our investigation of existing automated methods and, in particular, an analysis of their weaknesses has brought to light a number of ideals for a new automated method. These are outlined below and will guide the development of a new approach.

- Automated scores should be predictive of human judgements at the text level and, preferably, also at the sentence level;
- An ideal method should produce finely differentiated scores (rather than mere ranks) between competing translations, whether of very good or very poor quality;
- The method should produce reliable scores for MT *and* human translations;
- It should be informative for users *and* developers;
- The approach should be usable with different text types in different domains;
- It should be easily replicable;
- It should be adaptable for use with any language pair;
- Unlike methods such as BLEU, a new approach should aim to account for the level of quality of poorly translated sentence components (eg. by using weighting mechanisms) rather than assuming that all tokens in a text are of equal importance;
- Results should be comparable across evaluations.

3.5.7.2 Aims of this research

The aim of this research will be to devise a new automated method that meets as many of the above criteria as possible. However, in addition, we wish to explore the potential of an approach that does not require access to the source text or human reference translations. The non-requirement of human translations would enable the fast evaluation of any sample of output from any MT system.

Although some evaluation research, including work by (Callison-Burch and Flournoy, 2001), (Corston-Oliver et al., 2001), (Gamon et al., 2005) is based on the analysis of the MT output alone, correlations between automated scores/ranks and human judgements remain unexplored, as this was not the objective of this work. Furthermore, (Callison-Burch and Flournoy, 2001) found that only in 67%-94% of cases did the program automatically select the most fluent sentence, as rated by a human, from a set of

candidates. In response to this, it is our intention to explore whether a new method that relies solely on the MT output is capable of producing scores that reflect intuitive human judgements.

If an automated method is to be informative for users and developers, it should (a) produce reliable scores to enable users to compare competing outputs and (b) pinpoint and classify types of errors to help developers carry out system modifications. Such an automated tool could also be useful for post-editors, enabling the quick identification of items for revision, or the possibility of automatic correction in certain cases. An investigation of how errors in MT output can be identified, classified and automatically detected without the source text or a human translation is a previously unexplored area.

The automatic detection of errors will involve the use of some kind of test suite, designed to generate scores that correlate as highly as possible with human judgements. Following work by (Shiwen, 1993), a suite of algorithms will be developed to detect errors that typically occur in MT output for one language pair, with a view to extending the program to cover additional languages. In this case the target language will be English, and the initial aim will be to investigate the reliability of the algorithms with French as a source language before considering adaptability for other source languages.

The development of such a set of algorithms will first involve the manual analysis of output from different MT systems to identify and classify the various errors observed. The aim will then be to develop approaches to automatic error recognition based on human processing rationales, the intention being to automatically detect a number of different error types. This will enable experimentation with scores generated from different combinations of error categories, increasing the ability to find strong correlations with intuitive human judgements.

Chapter 4

Corpus compilation and annotation

Before exploring the potential of a new automated method for MT evaluation, human judgements of translation quality were first required, against which automated scores could later be compared. The two human evaluation metrics selected (see 3.4.10) determined the structure of the corpus to be used: for the evaluation of fluency and fidelity at the sentence level, requirements in terms of corpus content were as follows:

- texts in the chosen source language(s);
- English machine translations produced by the systems to be evaluated;
- An English human translation of each source text (to provide a benchmark against which scores for MT would be compared);
- one human reference translation[1] of each ST for the fidelity evaluation.

Some corpora for MT evaluation are publicly available and their use was considered for this work. However, these were found to be unsuitable for a number of reasons. The limitations of these corpora are presented in 4.1 along with a discussion on why the compilation of a new purpose-built corpus was essential for the current research.

The construction of a new corpus first required a detailed rationale for design (Elliott et al., 2003). This involved the consideration of specific text types for inclusion, the selection of language pairs to be evaluated, the number and length of texts to be incorporated, and the choice of MT systems to provide output for evaluation: these issues are discussed in 4.2. The compilation and structure of the corpus, including details of the MT systems from which the translations were generated, are described in 4.3.

[1] An expert translation, conveying all the factual information of the original without any stylistic flourishes. See 3.4.5.3.

The next stage involved the annotation of a sample of human and machine translations from the corpus with scores representing fluency and adequacy at the sentence level (Elliott et al., 2004a). The human evaluations, from which these scores were obtained, together with results and observations, are presented in 4.4.

The final stage of corpus preparation involved the manual annotation and classification of errors in all MT output for which human scores were available. This was conducted using the MT output alone, the ultimate aim being to automatically detect selected error types without access to the source text or any human translation. These annotations would provide a clear picture of the types of errors frequently found in French-English MT output and were intended to assist the selection of error types for automatic identification. Manually annotated errors would also enable us to evaluate the accuracy of our automated detection methods. The manual annotations and the resulting fluency error classification scheme (Elliott et al., 2004b) are described in 4.5. Concluding remarks on corpus compilation and annotation are presented in 4.6.

4.1 Existing corpora for MT evaluation

An investigation of MT evaluation methods in Chapter 3 has shown that many corpora, involving different language pairs, have been compiled for specific commercial and research-driven evaluation projects. However, these resources are either not publicly available or comprise text types or language pairs that do not suit our needs. For the initial stages of this research, a decision was taken to work with a corpus of French source texts with English human and machine translations. This language pair was selected for two reasons: (a) a large number of MT systems are available to translate from French into English and (b) both languages are known by the author, which is essential for corpus compilation (see 4.3).

Of the human and automated evaluations described in Chapter 3, only three French-English corpora were used, two of which are very small. These are summarised in Table 4.1.

Corpus	No. of French source texts	No. of words per ST (approx)	No. of machine translations	No. of human translations	No. of human reference translations	Available human scores
DARPA 1994	100 news texts	400	5	1	1	Fluency Fidelity Informativeness
Somers and Wild 2000	3 semi-technical texts	500	3	1	0	Cloze test
Callison-Burch and Flournoy 2001	Information not published	Information not published	2	0	0	Fluency

Table 4.1: Existing French-English corpora for MT evaluation

While the publicly available DARPA corpus (White et al., 1994) could be used for this research (it contains more than enough texts for our needs and includes human scores for fluency and adequacy at the sentence level), it contains only news texts, which do not reflect real use of MT. White states in (Somers, 2003) that "the purpose of declarative evaluation is to measure the ability of an MT system to handle text representative of actual end use", and our overview of the use of MT systems in 2.2 shows that they are not widely used for the translation of news articles: companies and other organisations tend to use MT to translate documents of a more technical nature. With this in mind, a decision was taken to identify the kinds of texts translated most frequently by MT systems to guide the compilation of a new corpus that reflects the needs of MT end-users.

4.2 Rationale for corpus design

In order to identify the kinds of texts translated most frequently by MT systems, a survey of MT users was conducted (Elliott et al., 2003). Responses to the survey provided information to guide the selection of texts for our corpus. Findings and decisions are presented in 4.2.1. Initial research focuses on the evaluation of French-English translations. However, it is also our intention to test error detection algorithms on other language pairs in which the target language is English. Consequently, the survey also provided information on languages frequently translated by MT systems to guide the selection of additional language pairs for this work. Findings are reported in 4.2.2. As the corpus was to be

compiled by the author alone, and the expense of obtaining human judgements had to be taken into consideration, careful thought was given to the size of the corpus. Experiments addressing corpus size and the validity of human scores are described in 4.2.3, enabling a decision to be taken on the number and length of texts for inclusion.

4.2.1 Text types

Since expectations of MT systems have become more realistic, many uses have been found for imperfect, raw MT output and users have identified texts to which MT is better suited. Various different text types, genres and subject matter are now machine-translated for different text-handling tasks, including filtering, gisting, categorising, information gathering and post-editing (White 2000). It is crucial, therefore, to compile a corpus for MT evaluation comprising texts that represent real use of MT.

A survey of MT users was conducted in January 2003 to gather information on the kinds of texts and subject matter most frequently translated by MT systems (Elliott et al., 2003). A short questionnaire was sent by email to a number of known MT users and researchers and to various mailing lists, including:

- aamt0002@infotokyo.ne.jp - Asia-Pacific Association for Machine Translation: mailing list for MT users, researchers and other interested parties;
- AMTAInfo@att.net - Association for Machine Translation in the Americas: mailing list for MT users, researchers and other interested parties;
- catmt@yahoogroups.com - for users, researchers and others interested in MT and computer-assisted translation systems;
- corpora@hd.uib.no - for academics and others interested in any area of corpus linguistics;
- linguist@listserv.linguistlist.org for academics and others interested in any area of language and linguistics;
- mt-list@eamt.org – European Association for Machine Translation: mailing list for MT users, researchers and other interested parties;

The questionnaire was intentionally concise, in an attempt to maximise the number of responses. Furthermore, the questions were provided in the body of the email, rather than in an attachment, to minimise the amount of time required to read the questions and to insert a response. The questionnaire is shown in Appendix 1. Despite the large number of recipients, and requests to forward the email to other MT users, only 25 responses were received[2]. Of these, 16 came from translation providers or international corporations and organisations, and the remaining 9 came from individual users of MT (mainly academics who needed to get the gist of research papers in unfamiliar languages). Four of the responses had to be discounted, as they came from users of computer-assisted translation tools, who had misunderstood the meaning of "fully-automatic MT systems". It was encouraging, however, that respondents included large companies and other institutions such as SMART Communications, Inc., SDL International, the European Commission, CLS Corporate Language Services, the Pan American Health Organisation, SAP AG and Microsoft Research. Full details of responses are shown in Appendix 2.

Recipients were asked to give details of the text types they translate with the aid of fully automatic MT systems, along with the approximate monthly word or page count, and specific subject domains. Three main problems were identified during the analysis of data:

1. Responses were clearly based on the MT users' own interpretations of text types and domains, and text categories provided by different respondents were found to overlap. There was also some variation in the level of detail provided. For example, some technical documents were listed merely as "technical", whereas others were placed in more specific categories, such as internal company documents, user manuals, instruction booklets, academic papers and web pages. This made text categorisation more complex than anticipated and had to be taken into account when selecting texts for our corpus.
2. Some respondents did not specify, or were unable to provide, specific information about the subject matter of the material they translate, and many were unable to give details of the number of words or pages translated per month. This prevented an accurate comparative analysis of the number of text types translated by large

[2] Such a low proportion of responses is not unusual, and has still allowed useful feedback for other Language Engineering projects, eg. (Atwell et al., 2000), (Al-Sulaiti and Atwell, 2004).

companies and organisations: some reported to translate thousands of words per month, whereas others who were likely to translate similar amounts could not provide this information;

3. It was difficult to equate the comparatively small number of words translated by single users with the thousands or millions of words translated by large companies and other organisations.

In response to (1) above, text types and subject domains were combined and categorised in such a way as to enable the analysis of all responses provided. With regard to (2) and (3), replies from companies/organisations and individual users were analysed separately, and the number of respondents translating each text type/domain was logged, rather than the total number of words, as this was impossible to calculate. Figure 4.1 shows text types translated by companies, organisations and professional translators and Figure 4.2 shows texts translated by individual users who use MT for personal, non-commercial translation (Elliott et al., 2003).

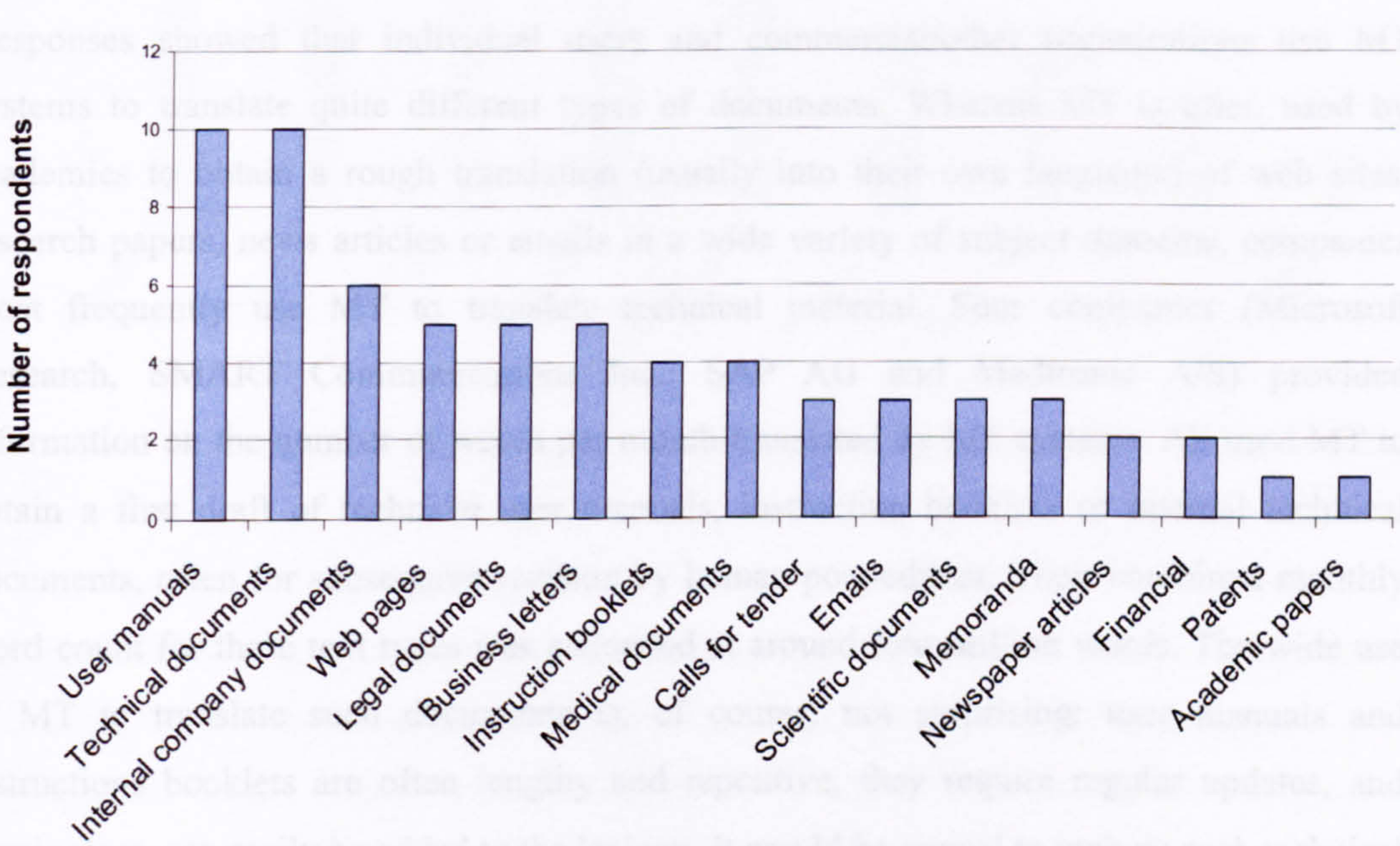


Figure 4.1: Text types machine-translated by companies, organisations and professional translators

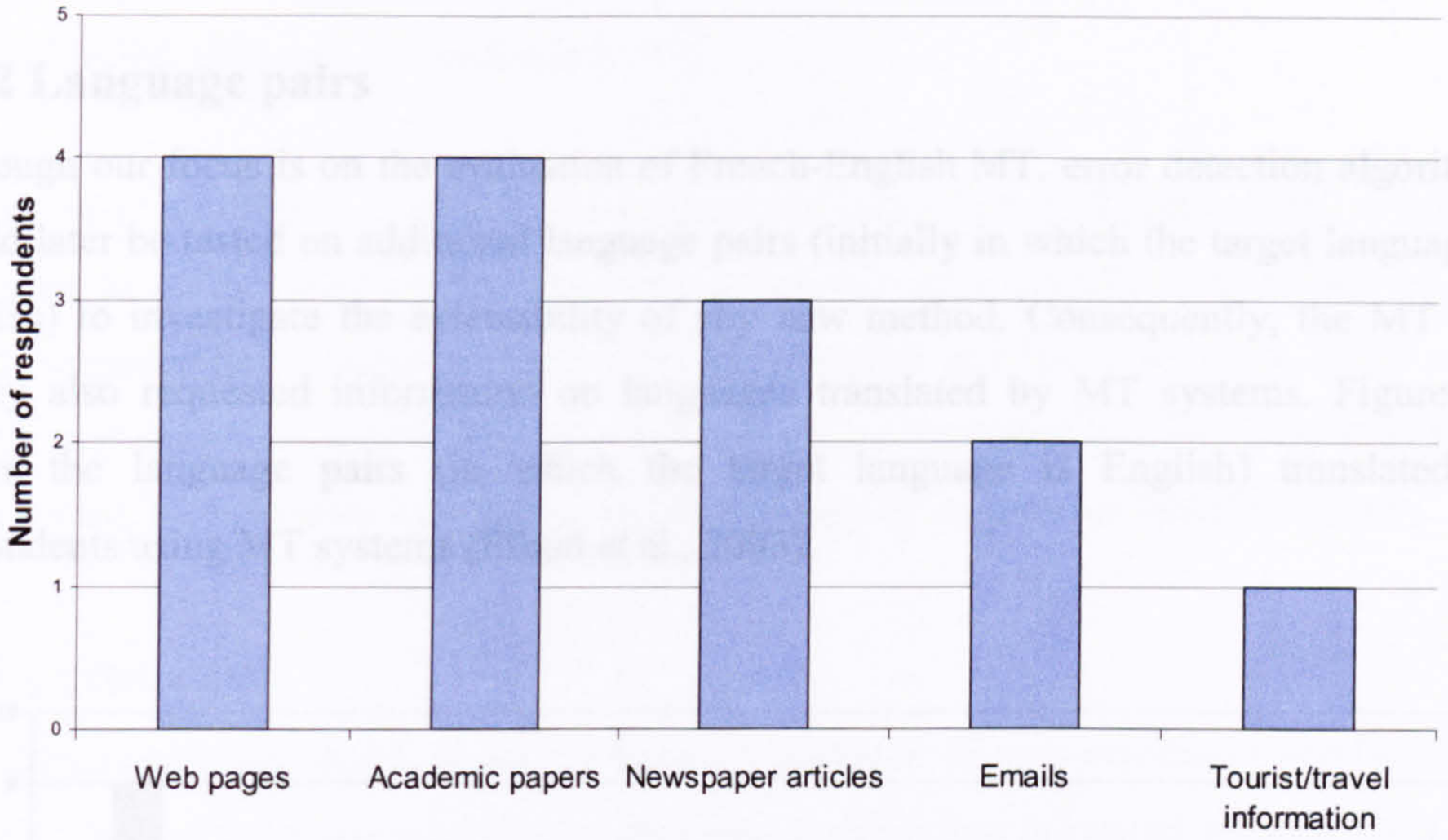


Figure 4.2: Text types machine-translated by individual users for personal, non-commercial use

Responses showed that individual users and commercial/other organisations use MT systems to translate quite different types of documents. Whereas MT is often used by academics to obtain a rough translation (usually into their own language) of web sites, research papers, news articles or emails in a wide variety of subject domains, companies most frequently use MT to translate technical material. Four companies (Microsoft Research, SMART Communications Inc., SAP AG and Medtronic A/S) provided information on the number of words per month translated by MT systems. All used MT to obtain a first draft of technical user manuals, instruction booklets or internal technical documents, often for subsequent revision by human post-editors. Their combined monthly word count for these text types was estimated at around four million words. The wide use of MT to translate such documents is, of course, not surprising: user manuals and instructions booklets are often lengthy and repetitive, they require regular updates, and terminology can easily be added to the lexicon. It would be crucial to include such technical documents in our corpus to reflect the commercial use of MT. Texts selected for the corpus are described in 4.3.1.

4.2.2 Language pairs

Although our focus is on the evaluation of French-English MT, error detection algorithms would later be tested on additional language pairs (initially in which the target language is English) to investigate the extensibility of any new method. Consequently, the MT user survey also requested information on languages translated by MT systems. Figure 4.3 shows the language pairs (in which the target language is English) translated by respondents using MT systems (Elliott et al., 2003).

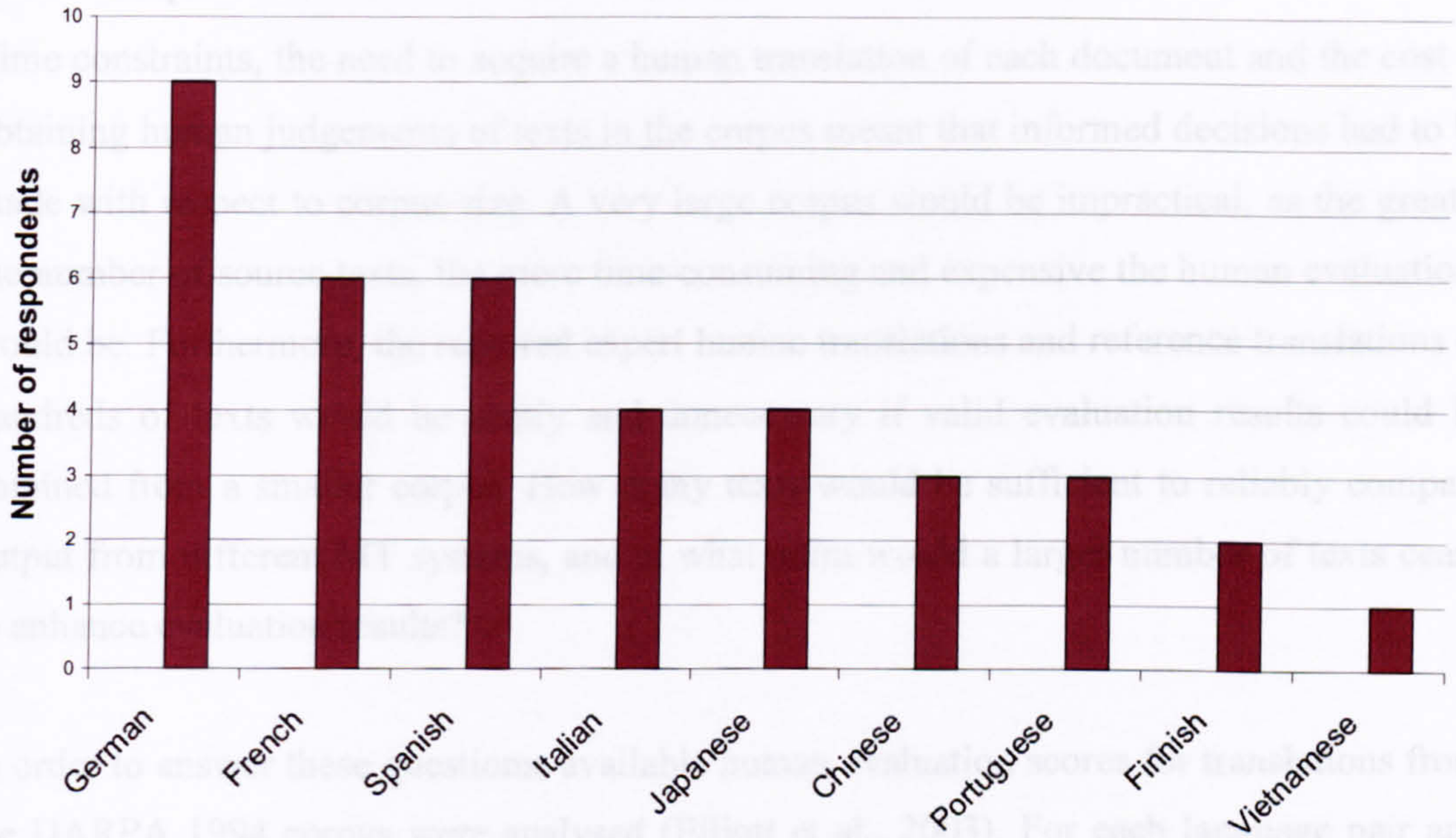


Figure 4.3: Languages machine translated into English

The number of language pairs that MT systems are now able to handle is constantly increasing. The International Association for Machine Translation (IAMT) Compendium of Translation Software (Hutchins and Hartman, 2005) lists a large number of MT systems that translate between many more languages than those shown in Figure 4.3. Yet importantly, the five source languages most frequently translated into English according to our survey were found to coincide with the top five source languages in the Compendium:

German-English (49 available MT systems), Japanese-English (44), Spanish-English (43), French-English (38) and Italian-English (25).

In view of these findings, after initial evaluations using French-English texts, a decision was taken to collect additional source texts to test our error detection algorithms on three further language pairs: German-English, Spanish-English and Italian-English. These reflect the MT user responses shown in Figure 4.3 and can be translated by MT systems available to the author. Experiments with Japanese-English may be an area for further research.

4.2.3 Corpus size

Time constraints, the need to acquire a human translation of each document and the cost of obtaining human judgements of texts in the corpus meant that informed decisions had to be made with respect to corpus size. A very large corpus would be impractical, as the greater the number of source texts, the more time-consuming and expensive the human evaluations would be. Furthermore, the required expert human translations and reference translations of hundreds of texts would be costly and unnecessary if valid evaluation results could be obtained from a smaller corpus. How many texts would be sufficient to reliably compare output from different MT systems, and at what point would a larger number of texts cease to enhance evaluation results?

In order to answer these questions, available human evaluation scores for translations from the DARPA 1994 corpus were analysed (Elliott et al., 2003). For each language pair and attribute (adequacy, fluency and informativeness) a mean score was computed for each possible number of translations (between one and one hundred) by each MT system. Figures 4.4, 4.5 and 4.6 show the mean score by system for each number of French-English texts evaluated for each of the three attributes, each text comprising around 400 words (Elliott et al., 2003).

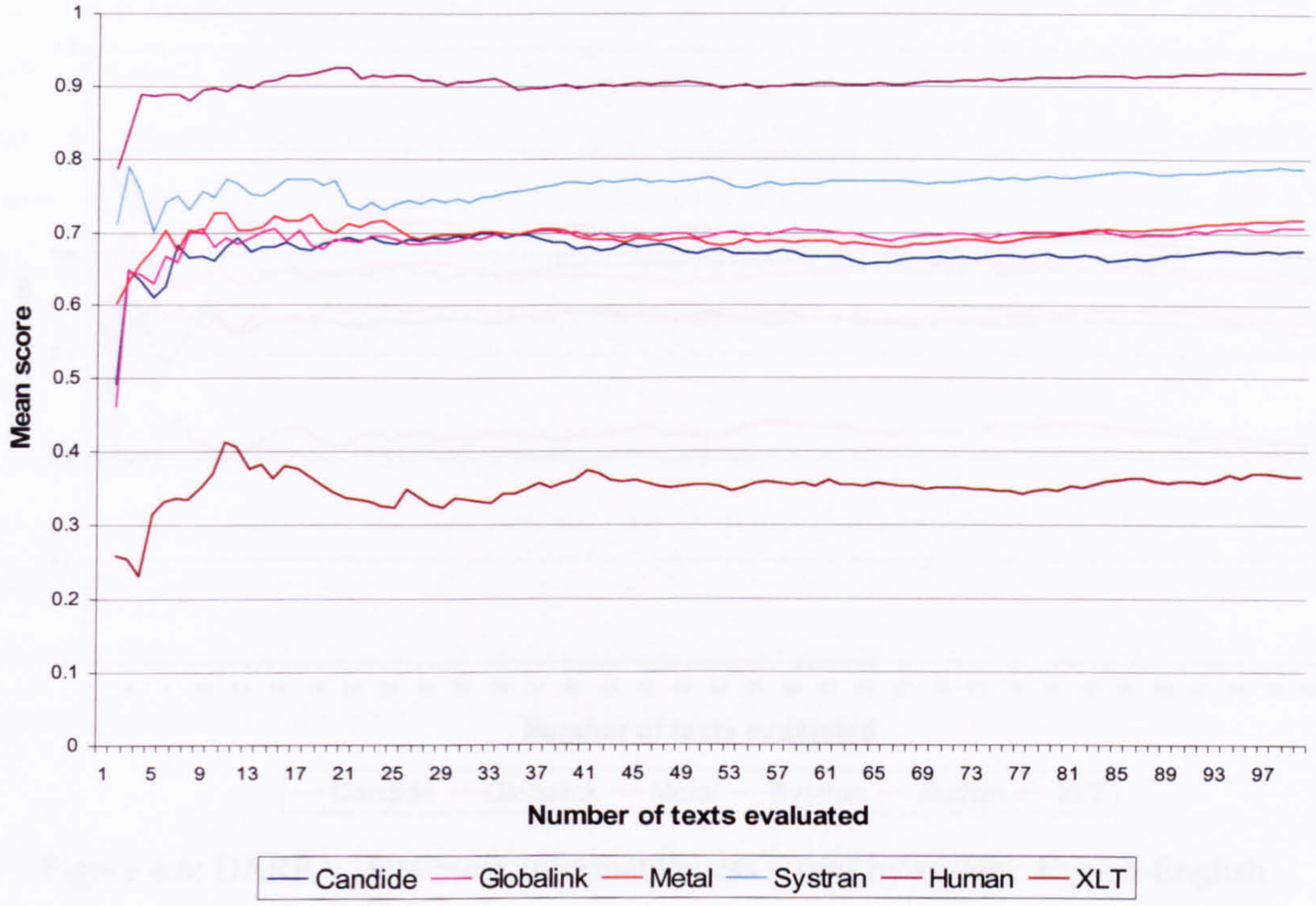


Figure 4.4: DARPA 1994 mean adequacy scores by system: French-English

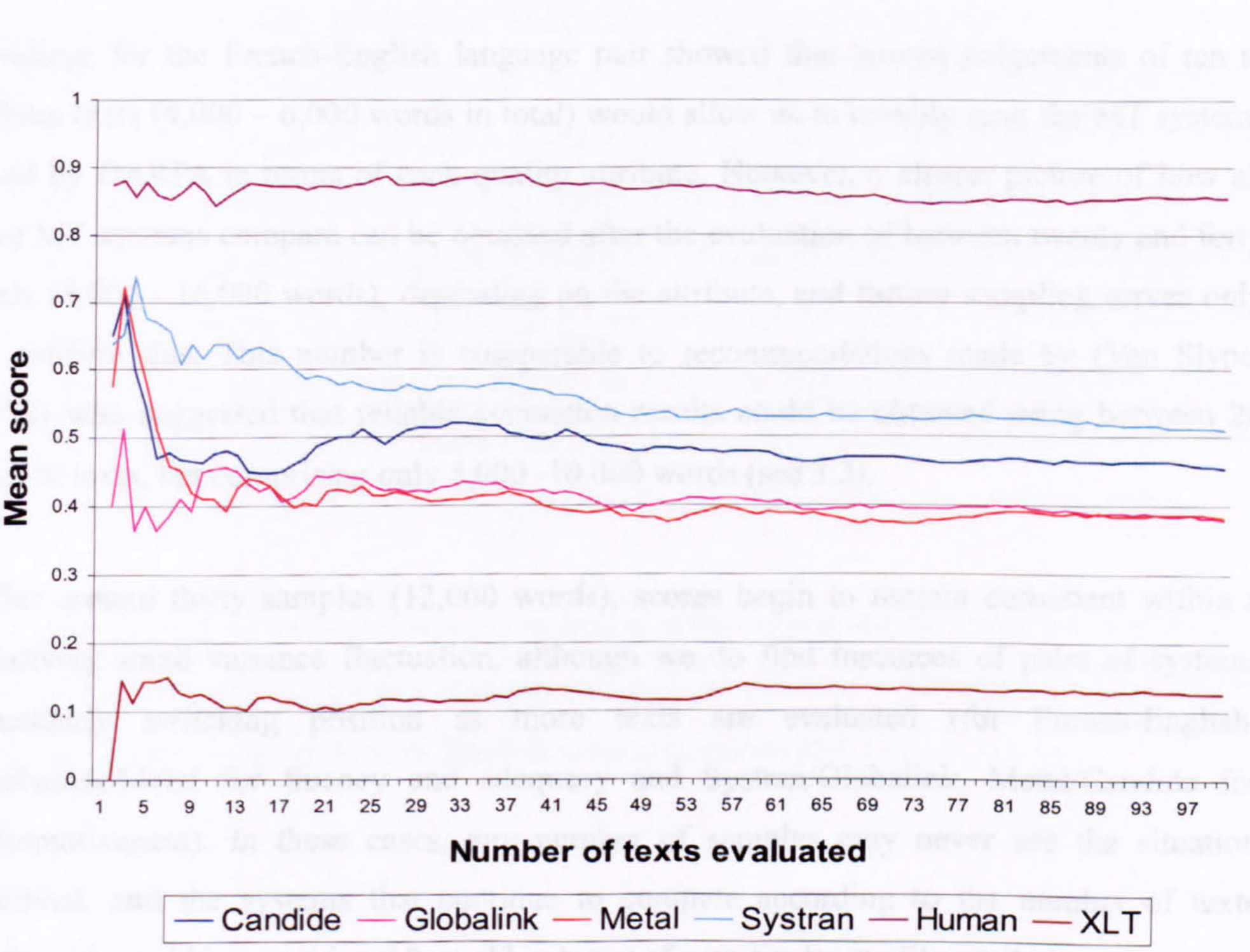


Figure 4.5: DARPA 1994 mean fluency scores by system: French-English

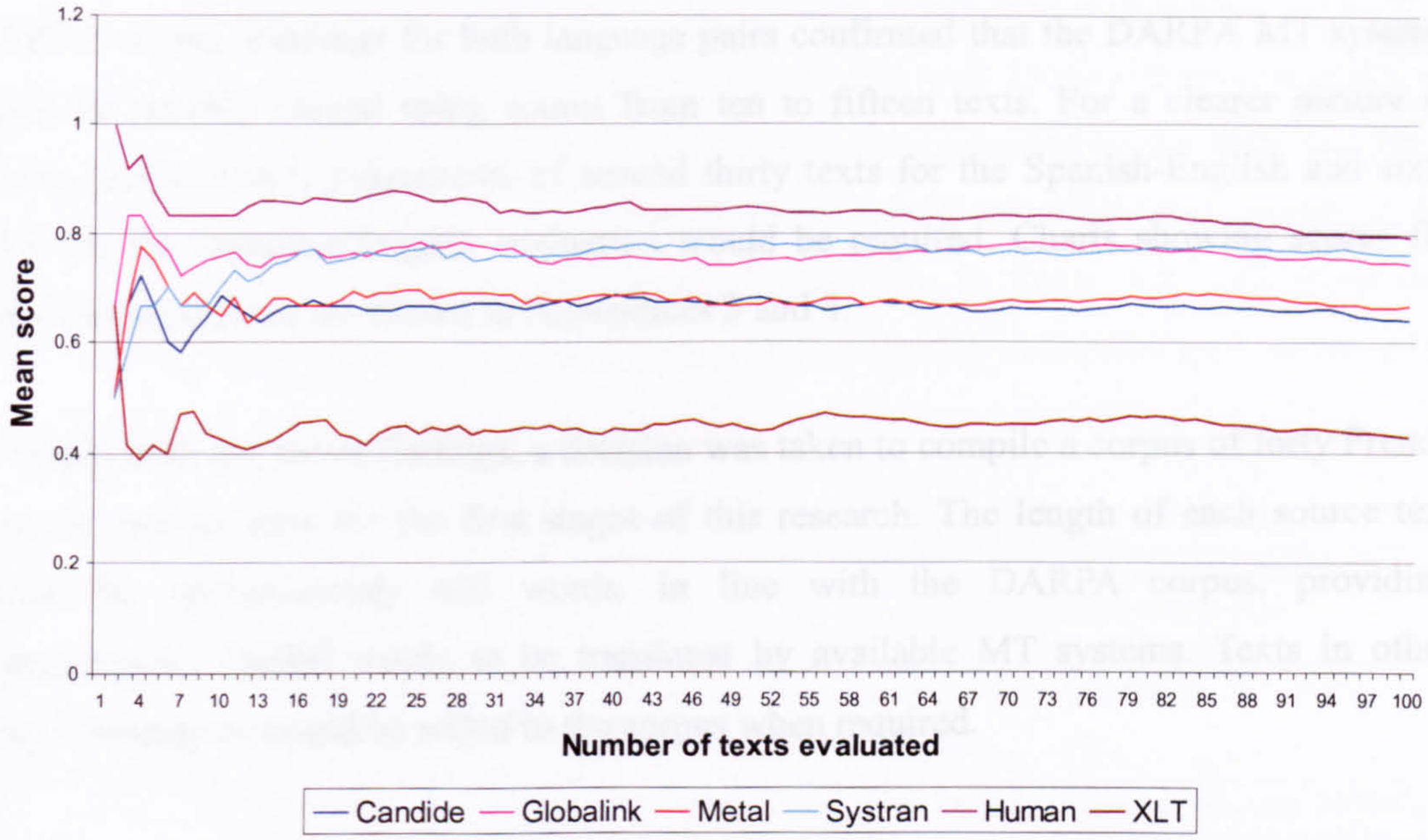


Figure 4.6: DARPA 1994 mean informativeness scores by system: French-English

Findings for the French-English language pair showed that human judgements of ten to fifteen texts (4,000 – 6,000 words in total) would allow us to reliably rank the MT systems used by DARPA in terms of each quality attribute. However, a clearer picture of how all five MT systems compare can be obtained after the evaluation of between twenty and forty texts (8,000 - 16,000 words), depending on the attribute, and further sampling serves only to confirm this. This number is comparable to recommendations made by (Van Slype, 1979) who suggested that reliable evaluation results could be obtained using between 20 and 40 texts, but comprising only 5,000 -10,000 words (see 3.3).

After around thirty samples (12,000 words), scores begin to remain consistent within a relatively small variance fluctuation, although we do find instances of pairs of systems constantly switching position as more texts are evaluated (for French-English: Globalink/Metal for fluency and adequacy and Systran/Globalink, Metal/Candide for informativeness). In these cases, any number of samples may never see the situation resolved, and the systems that continue to compete according to the number of texts evaluated, could be considered "equal" in terms of a particular quality attribute.

The same analysis was carried out using the Spanish-English and Japanese-English DARPA scores. Findings for both language pairs confirmed that the DARPA MT systems could be reliably ranked using scores from ten to fifteen texts. For a clearer picture of system performance, judgements of around thirty texts for the Spanish-English and sixty texts for the Japanese-English evaluation would be required. Charts showing scores for these language pairs are shown in Appendices 3 and 4.

In response to the above findings, a decision was taken to compile a corpus of forty French texts to provide data for the first stages of this research. The length of each source text would be approximately 400 words, in line with the DARPA corpus, providing approximately 16,000 words to be translated by available MT systems. Texts in other source languages would be added to the corpus when required.

4.3 Corpus compilation and structure

As discussed in 4.2.1, our intention was to collect French source texts that reflect commercial use of MT. In response to findings from the MT user survey, a decision was taken to compile a corpus of technical texts in the computing domain; representing the two text types most frequently machine translated by respondents (see Figure 4.1): (1) user manuals and (2) other technical documents. Our aim was to select ten extracts from software user manuals, ten extracts from FAQs (frequently asked questions) on software applications (enabling the evaluation of the chosen systems' ability to translate questions), ten extracts from technical reports relating to software applications and ten press releases on software topics. This would enable the representation of a greater variety of verb tenses, as user manuals and FAQs tend to contain mostly imperative and present tense verbs.

The identification of suitable texts for inclusion in the corpus and the problems encountered are discussed in 4.3.1. The MT systems selected for evaluation are listed in 4.3.2 and the structure of the complete French-English corpus is shown in 4.3.3.

4.3.1 Text selection: problems and solutions

The first phase involved an investigation into existing freely-available parallel corpora to determine whether any appropriate original French texts with English human translations could be selected. Multilingual and bilingual corpora were explored on the Web, and those containing texts inappropriate for our needs were immediately discarded. Some of the corpora investigated further are displayed alphabetically in Table 4.2. The comments illustrate the kinds of difficulties encountered while attempting to identify appropriate texts. (NB. Comments reflect findings in 2003.)

Corpus and location	Comments
BAF (Bi-text anglais-français) French-English/English-French parallel corpus http://rali.iro.umontreal.ca/	Freely available. Some technical reports related to software applications are usable
ECI/MCI 1 European Corpus Initiative Multilingual Corpus http://www.elsnet.org/resources/eciCorpus.html	Not freely available
ET10-63 Parallel Corpus http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html	EC documents on telecoms. Possible texts of relevance, but not available to download online
ITU (International Telecommunications Union) CRATER Parallel Spanish-French-English corpus http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html	Some texts related to software but not ideal – many comprise long lists and incomplete sentences
Knut Hofland English-French Aligned texts http://kh.hd.uib.no/tactweb/en-fr.htm	Database search facility out of operation
MULTEXT (Multilingual Text and Corpora) http://www.lpl.univ-aix.fr/projects/multext/	Limited amount of material for download is inappropriate for our corpus
TELRI: Trans European Language Resources Infrastructure: TRACTOR archive http://www.ids-mannheim.de/telri/cdrom.html	Multilingual parallel texts. Text types not viewable without subscription
United Nations Parallel Text Corpus http://morph.ldc.upenn.edu/Catalog/LDC94T4A.html	LDC (Linguistic Data Consortium) membership required.

Table 4.2: Corpora investigated for appropriate texts

Having identified some suitable material from the online BAF corpus (see Table 4.2), the second phase involved a general search for appropriate French texts with existing English translations on the Web. First, the use of the STRAND system (Resnik 1998, 1999, 2003) was considered for automatically mining parallel texts from the Web. However, using such a method would (a) return texts in any domain, many of which would be inappropriate for our needs; (b) provide us with translations varying widely in terms of quality and (c) would inevitably return texts that were not strictly parallel: in particular, high quality translations may not correspond at the sentence level with the original; a professional translator may,

for instance, insert additional information for the needs of the target audience, omit information considered irrelevant to the target audience or split/combine sentences to produce a more easily readable translation. Any of these changes would hinder segment correspondence for our evaluations. For these reasons, it was decided to manually locate appropriate texts and to ensure that they were parallel at the sentence level.

Web searches were performed to locate French user manuals, FAQs, technical reports and press releases in the software domain with existing English translations. However, finding French texts with good quality translations proved to be a difficult task. Some original texts (particularly user manuals and FAQs) were badly written, making them unsuitable as input to an MT system, whose performance can only be fairly assessed based on well-formed input. In many cases the translations were of very poor quality and some had obviously been written by native speakers of the source language. Other translations, although of excellent quality, were so far removed from the original that they were deemed unusable for MT evaluation.

In addition to considerations of language quality, texts also had to be understandable to regular users of computer applications, enabling evaluators to confidently judge the quality of the translations. Some texts were considered too technical for human evaluators to judge reliably, as they contained large amounts of terminology and phraseology that would be too difficult for non-specialists to understand. Furthermore, some texts were unsuitable as they comprised long lists of product codes and descriptions, often nothing more than noun strings and numbers. Such texts would prevent the evaluation of many grammatical features as they did not contain full sentences.

The INDEXA[3] website, a French business directory, proved to be a useful resource for appropriate texts, providing links to software companies, many of whose websites contained parallel pages in English. Potentially useful texts were checked for quality, according to the requirements mentioned above, in both the source and target languages before being selected or rejected. Permission to use selected texts was requested by email, as the intention is to make the corpus available to others for research. However, obtaining

[3] http://www.indexa.fr/

copyright permission was an arduous task, so methods were devised to locate suitable documents with a notice giving permission to use, copy, distribute or modify the text and/or its translations for non-commercial use. Searches in French and English for "Guide de l'utilisateur" (User Guide) + "reproduction permitted" and "logiciel libre" (free software) + "copyleft" (the term used by the Free Software Foundation to describe their uncopyrighted software programs and documentation) yielded useful results. Many texts produced under the GNU Free (software) Documentation Licence and by the Free Software Foundation Europe were selected for the corpus. Another fruitful method involved searching sites in France for relevant expressions in English (eg. User Guide English version, Copyright permitted).[4]

Having selected forty suitable texts, from a variety of different authors, companies and organisations, all source documents and their translations were meticulously checked for errors and translation correspondence. A large number of corrections were made, as only perfect MT input and good quality human translations would provide reliable evaluation results. Some translations, for instance, were not entirely aligned at the sentence level, some did not convey all of the information in the original and some contained grammatical or punctuation errors. Real MT input is rarely perfect, but removing imperfections would enable the evaluation of the true performance of each system; investigating the robustness of MT engines in the face of ill-formed input is another area for research. Several weeks later, without reference to the existing human translations, English reference translations, containing all of the information conveyed in the original but without stylistic flourishes, were produced for the human fidelity evaluations.

4.3.2 MT systems

Four rule-based MT systems were selected to translate the original French texts into English. Three of these were commercial systems available at the University of Leeds: Systran Translation Project Manager 4.0, Reverso ProMT 5.01 and Comprendium 1.0. In addition, SDL's FreeTranslation[5] ,was selected, as this was deemed by the author to offer the best French-English translation quality among free online systems. Babel Fish

---

[4] My thanks to Jeff Allen for this suggestion
[5] http://mobile.freetranslation.com/c/

Translation[6], Google Translate[7], Voilà[8] and Worldlingo[9] were not considered as these were all powered by Systran and produced almost identical translations to those generated by our own commercial version of Systran. No data-driven systems were available for our use at the time, but the evaluation of such systems is an area for future work.

4.3.3 Corpus structure

The forty source documents were translated by each of the four MT systems to provide all texts required for the initial stages of this research. The final version of the French-English corpus is summarised in Figure 4.7., comprising a total of approximately 112,000 words (Elliott et al., 2004a).

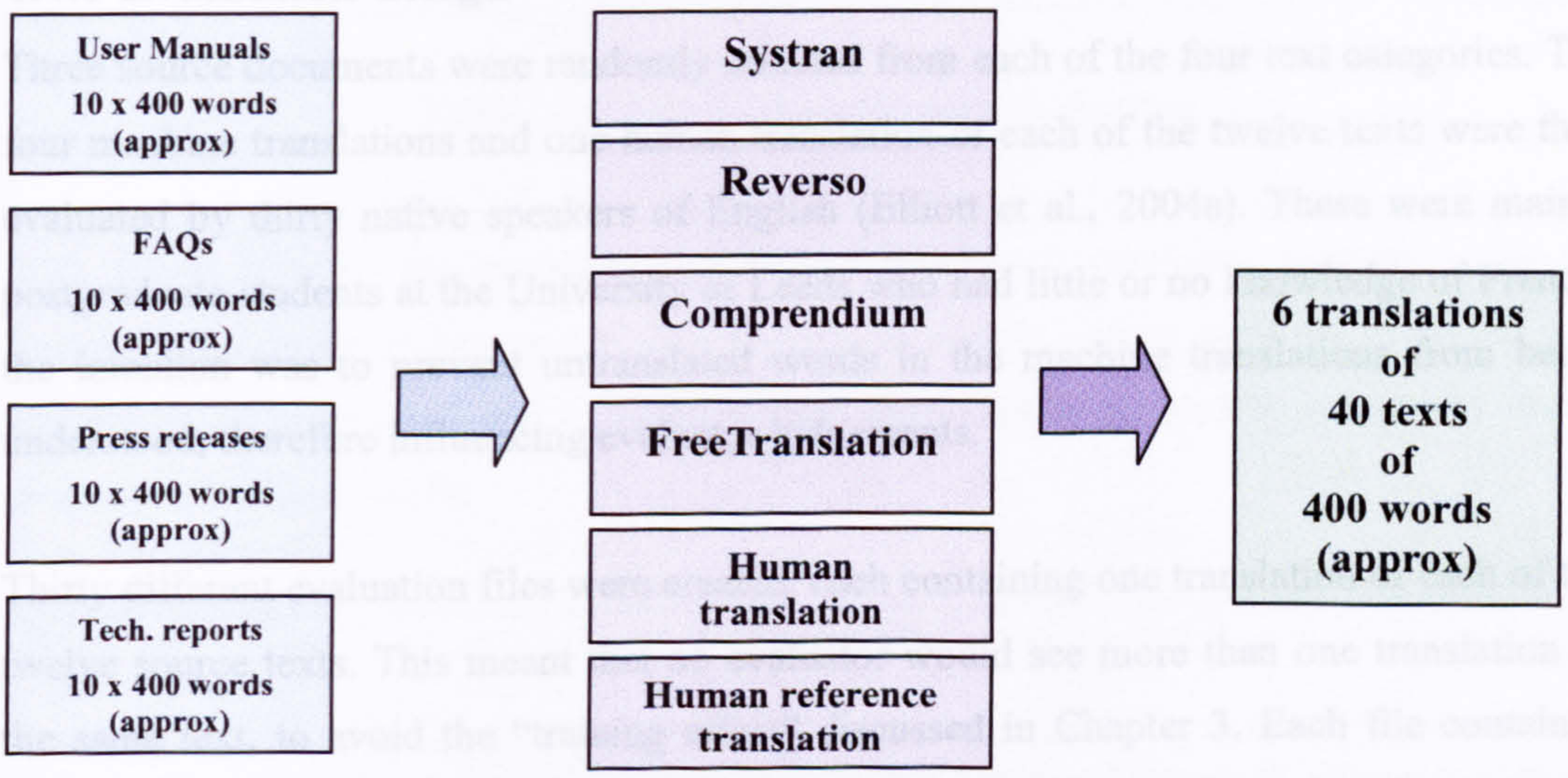


Figure 4.7: French-English corpus structure

4.4 Annotating the corpus with human judgements

The initial intention was to obtain human judgements of fluency and fidelity at the sentence level for the four machine translations and one candidate human translation of all forty source texts. However, due to financial limitations, a decision was taken to initially

---

[6] http://world.altavista.com/
[7] http://www.google.com/translate_t
[8] http://tr.voila.fr/
[9] http://www.worldlingo.com/

evaluate translations of a sample of twelve texts from the corpus, comprising three from each of the four text categories. The analysis of required corpus size described in 4.2.3 indicated that twelve texts should still provide reliable results. This in itself would prove costly as multiple human judgements for fluency and fidelity would be required for each translation due to the subjective nature of these methods (see 3.4.1 and 3.4.5). The remaining 28 texts in the corpus are still useful for further work on this project.

Practical issues relating to the design of the evaluations are presented in 4.4.1. The fluency and fidelity evaluations are then described in 4.4.2 and 4.4.3, time required for the evaluations is discussed in 4.4.4 and results and observations are presented in 4.4.5.

4.4.1 Evaluation design

Three source documents were randomly selected from each of the four text categories. The four machine translations and one human translation of each of the twelve texts were then evaluated by thirty native speakers of English (Elliott et al., 2004a). These were mainly postgraduate students at the University of Leeds who had little or no knowledge of French; the intention was to prevent untranslated words in the machine translations from being understood, therefore influencing evaluator judgements.

Thirty different evaluation files were created, each containing one translation of each of the twelve source texts. This meant that no evaluator would see more than one translation of the same text, to avoid the "training effect" discussed in Chapter 3. Each file contained translations from a different combination of sources (MT and human). Each evaluator rated six of the translations for fluency and six for adequacy (based on the measure of fidelity used by DARPA). This would provide three judgements per segment for each attribute, from which a mean score could be computed. In half of the files, the six fluency evaluations appeared first; the other half began with the adequacy evaluations. Judges were not told that the texts were translations. Subjects read the candidate sentences on a computer screen and entered their judgements electronically to facilitate the collation of scores and to avoid transcription errors.

4.4.2 Evaluation of fluency

The chosen approach was based on the DARPA 1994 fluency evaluation (White et al., 1994) described in 3.4.1.3. Texts were divided into fragments (ie. sentences or headings) and presented in the form of a table with a scoring box to the right of each 'candidate text', as shown in Figure 4.8.

No.	Candidate segment	Score
1.	6.2. Parental control	
2.	In the main menu select *Parameters, parental Control.*	
3.	In the list of computers select the wished computer.	
4.	It is necessary that the computer is that is authorized, marked, to define a filter.	

Figure 4.8: Extract from Fluency Evaluation

With access only to the translation, evaluators rated each candidate segment using the fluency scale shown in Figure 4.9. To simplify the metric, judges were not provided with definitions for scores 2, 3 and 4. Subjects were asked not to go back to a segment once a judgement had been made.

Fluency scale

Look carefully at each segment and give each one a score according to how much you think the text reads like fluent English written by a native speaker. Give each segment of text a score of 1, 2, 3, 4, or 5 where:

5 = All of the segment reads like fluent English written by a native speaker

1 = None of the segment reads like fluent English written by a native speaker

Figure 4.9: Fluency scale

4.4.3 Evaluation of adequacy

This approach was based on the DARPA 1994 adequacy evaluation (White et al., 1994) described in 3.4.5.3. The segmented translations were presented in the form of a table in

which each 'candidate text' was aligned with the human reference translation, as shown in Figure 4.10.

No	Reference text	Candidate text	Score
1	6.2. Parental Control	6.2. Parental control	
2	At the main menu select *Settings, Parental Control.*	In the main menu select *Parameters, parental Control.*	
3	Select the required computer from the list.	In the list of computers select the wished computer.	
4	The computer must be authorized (ticked) to define a filter.	It is necessary that the computer is that is authorized, marked, to define a filter.	

Figure 4.10: Extract from Adequacy Evaluation

Evaluators compared the candidate segment with the aligned 'reference text' and used the adequacy scale shown in Figure 4.11 to score each segment. As with the fluency evaluation, judges were not provided with definitions for scores 2, 3 and 4. Subjects were asked not to go back to a segment once a judgement had been made.

Adequacy scale

For each segment, read carefully the reference text on the left. Then judge how much of the same *content* you can find in the candidate text, *regardless of grammatical errors, spelling errors, inelegant style or the use of* synonyms. Give each segment of text a score of 1, 2, 3, 4, or 5 where:

5 = All of the content in the reference text is present in the candidate text

1 = None of the content is present (OR the text completely contradicts the information given on the left hand side)

Figure 4.11: Adequacy scale

4.4.4 Evaluation time required

Each evaluator judged 327 segments, rating approximately half for adequacy and half for fluency. The average time taken to complete the fluency evaluation was 33 minutes. The adequacy evaluation contained more reading material and took 48 minutes on average to

complete. Without including an introduction to the task, time needed to read instructions, and at least one break, 30 evaluators each required 81 minutes to complete the evaluations. Therefore, the total time needed to evaluate five translations of twelve texts amounted to 40.5 hours (Elliott et al., 2004a).

4.4.5 Results and observations

Scores from three evaluators for both fluency and adequacy were obtained for each translation of each segment. A mean score was then calculated per segment for each quality attribute and, subsequently, for each text, text type and system. Findings from the human evaluations are reported in two parts: system scores for fluency and adequacy are presented in 4.4.5.1 and scores according to text type are discussed in 4.4.5.2.

4.4.5.1 Evaluation results by system

Figure 4.12 and Table 4.3 summarise the human evaluation scores obtained for fluency and adequacy for the MT systems and human translations (Elliott et al., 2004a).

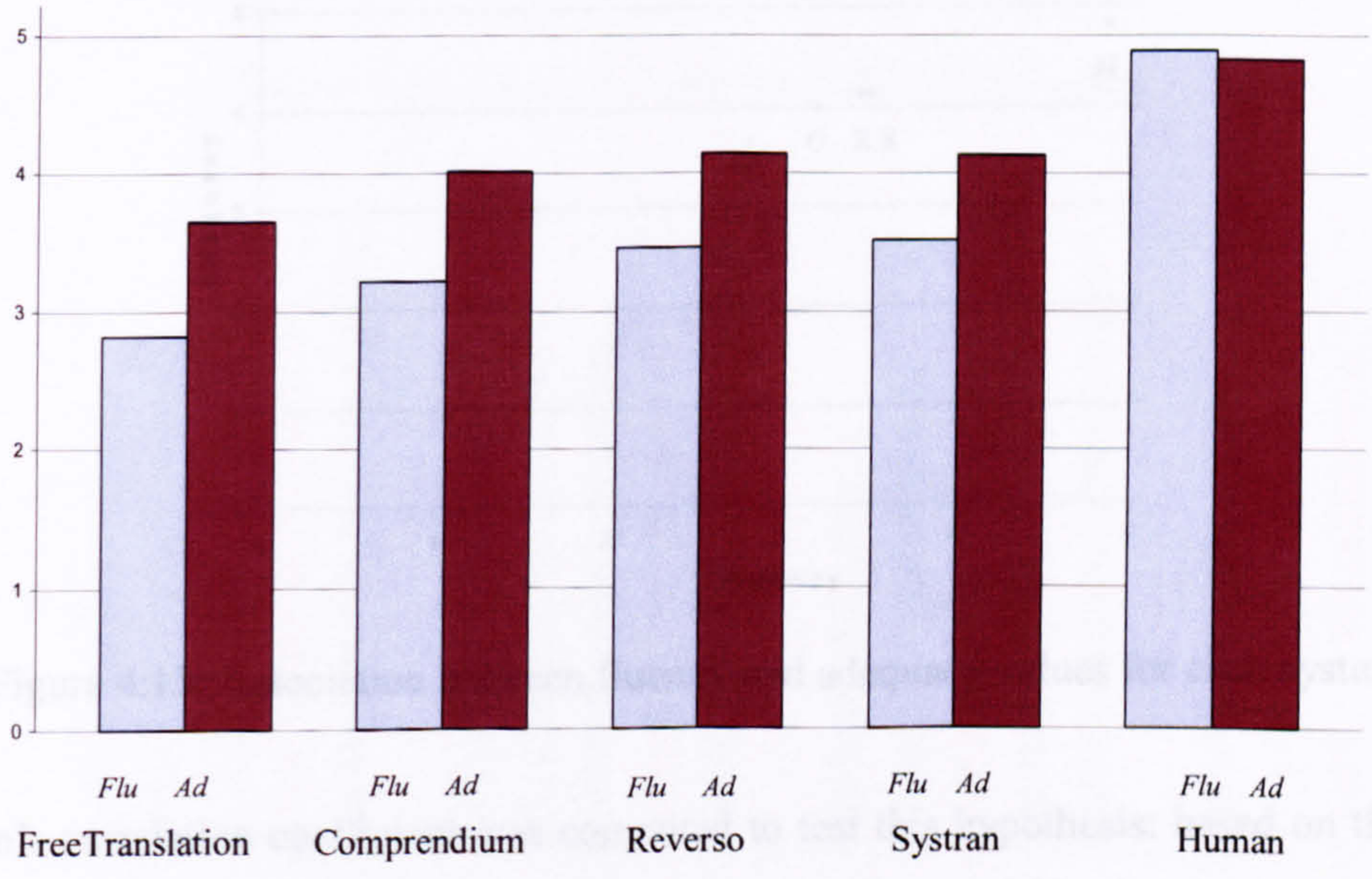


Figure 4.12: Mean scores for MT systems and human translations

Results show that Systran was the highest scoring MT system for fluency and Reverso for adequacy, each by a very small margin. Both systems outperformed Comprendium, and

FreeTranslation was the lowest scoring system for both attributes. As Figure 4.12 and Table 4.3 show, all MT systems scored more highly for adequacy than for fluency, indicating that despite a lower level of fluency, the content of raw MT output can be useful. Conversely, the human translations scored more highly for fluency, although the difference between fluency and adequacy scores (0.067) was negligible.

System	Fluency Score and Rank	Adequacy Score and Rank
Human	4.893 (1)	4.826 (1)
Systran	3.519 (2)	4.136 (3)
Reverso	3.466 (3)	4.142 (2)
Comprendium	3.221 (4)	4.013 (4)
FreeTranslation	2.827 (5)	3.644 (5)

Table 4.3: Mean Scores by System

For all five 'systems', a high degree of association was found between values for fluency and adequacy, as shown in Figure 4.13.

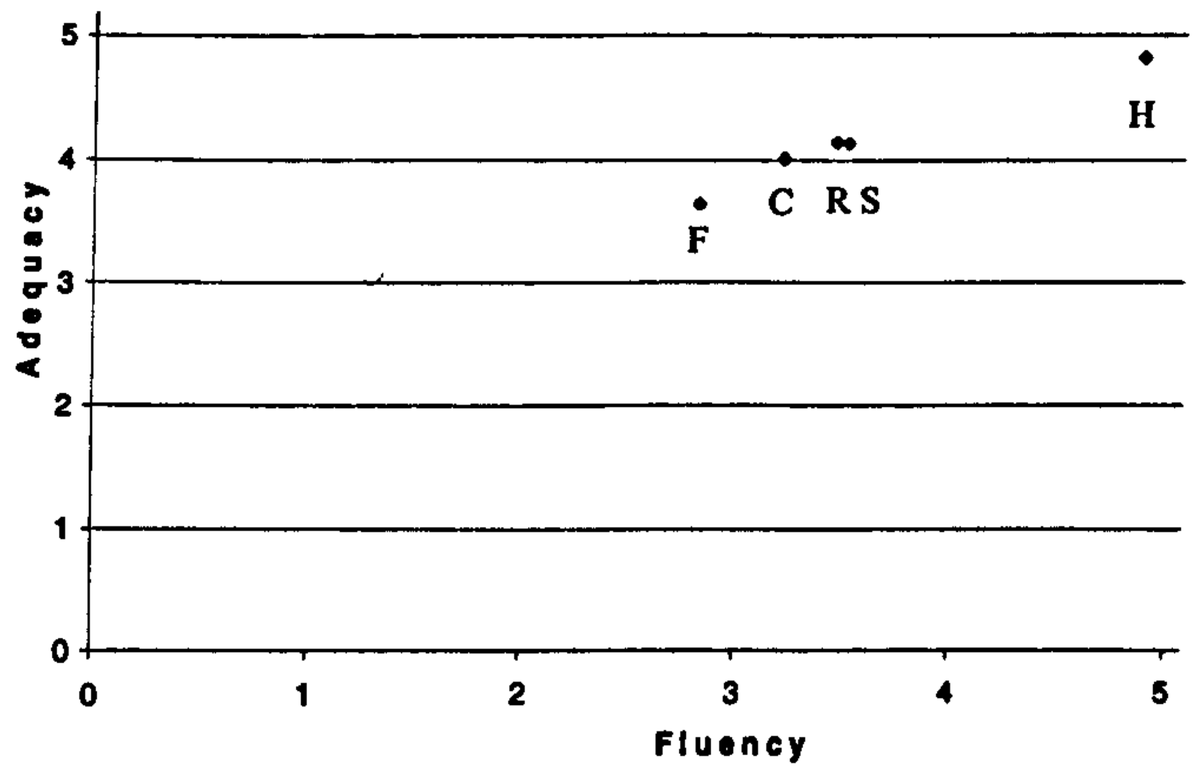


Figure 4.13: Association between fluency and adequacy values for each system

Pearson's correlation coefficient was computed to test this hypothesis: based on the mean system scores for fluency and adequacy, the value of $r = 0.988$, showing a very strong correlation between the two variables. This indicates that evaluating either fluency or adequacy would be sufficient to predict values for the other attribute and supports earlier findings such as those reported in Chapter 3 (eg. Carroll, 1966; White et al., 1994; White, 2001).

4.4.5.2 Evaluation results by system and text type

Differences between scores for each text type were subsequently investigated. Mean scores by text type and system for fluency and adequacy are shown in Figures 4.14 and 4.15.

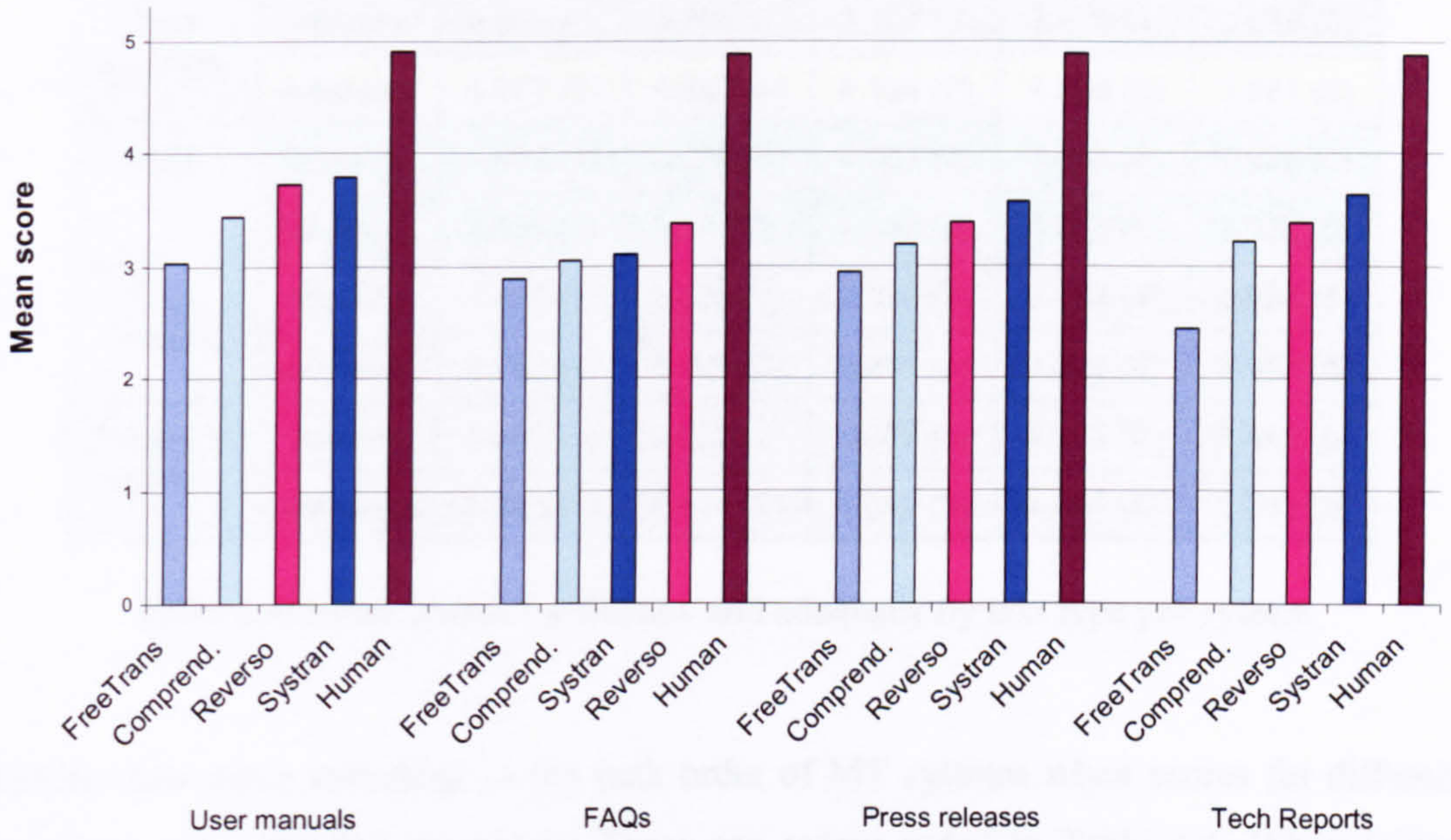


Figure 4.14: Fluency scores per text type by system

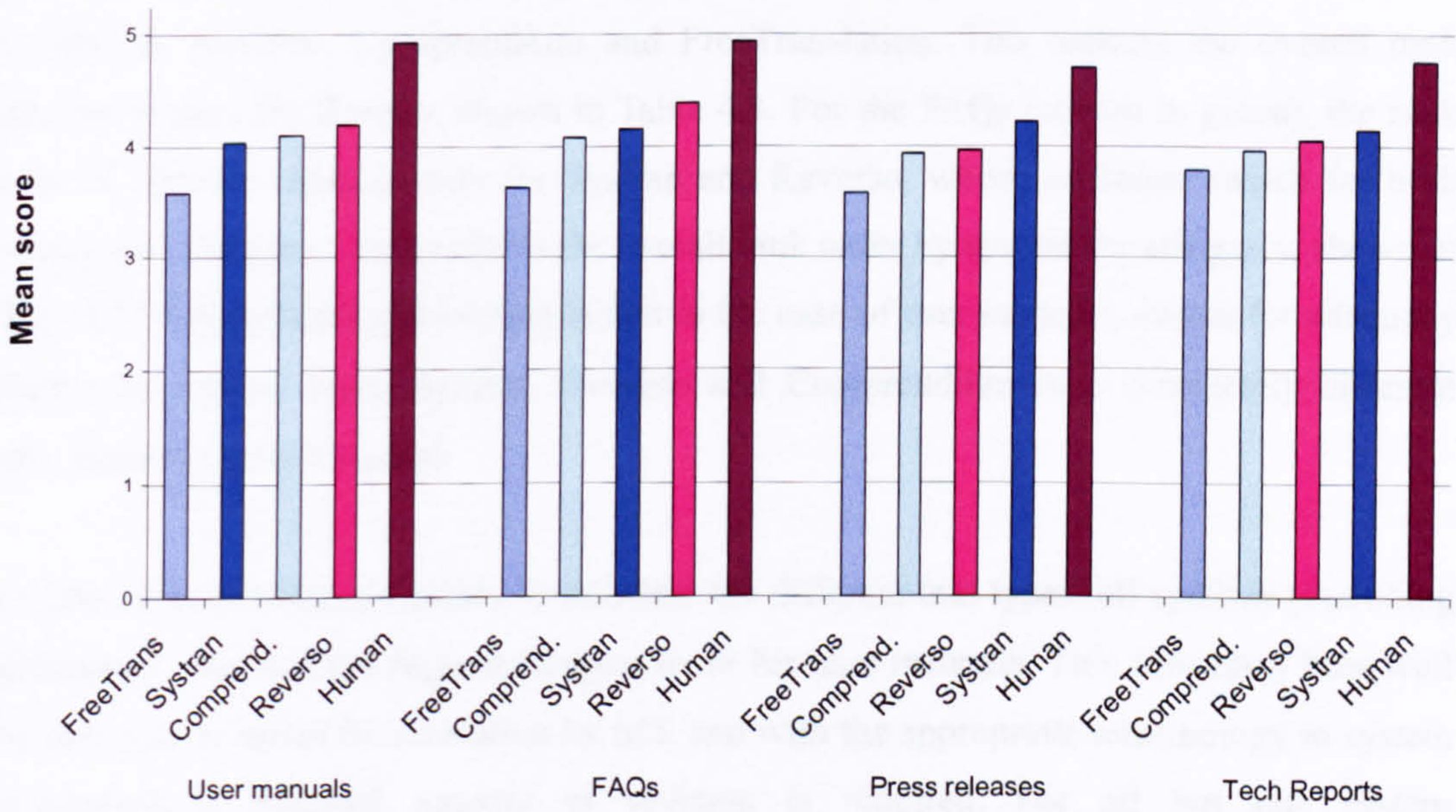


Figure 4.15: Adequacy scores per text type by system

Human scores for both attributes are presented together for comparison in Table 4.4 (Elliott et al., 2004a).

Text type	Attribute	Human	Systran	Reverso	Comprend	FreeTrans
User manuals	**Fluency**	4.913 (1)	3.783 (2)	3.727 (3)	3.438 (4)	3.036 (5)
	Adequacy	4.917 (1)	4.031 (4)	4.194 (2)	4.098 (3)	3.581 (5)
FAQs	**Fluency**	4.878 (1)	3.095 (3)	3.367 (2)	3.039 (4)	2.892 (5)
	Adequacy	4.933 (1)	4.147 (3)	4.380 (2)	4.074 (4)	3.628 (5)
Press releases	**Fluency**	4.896 (1)	3.563 (2)	3.390 (3)	3.194 (4)	2.935 (5)
	Adequacy	4.705 (1)	4.221 (2)	3.951 (3)	3.925 (4)	3.583 (5)
Technical reports	**Fluency**	4.887 (1)	3.633 (2)	3.382 (3)	3.213 (4)	2.445 (5)
	Adequacy	4.750 (1)	4.144 (2)	4.043 (3)	3.954 (4)	3.783 (5)

Table 4.4: Mean scores for fluency and adequacy by text type per system

Results show three variations in the rank order of MT systems when scores for different text types are computed separately. These are colour-coded in Table 4.4. When taking results for either fluency or adequacy into consideration, the most frequent rank order (shown in blue), according to text type, places the human translation in first place, followed by Systran, Reverso, Comprendium and FreeTranslation. This reflects the overall rank order by system for fluency, shown in Table 4.3. For the FAQs (shown in green), the rank order of systems changes only for Systran and Reverso, whose positions switch for both fluency and adequacy. This reflects the overall rank order by system for adequacy, shown in Table 4.3. A significant observation is that in the case of user manuals, scores for adequacy (shown in yellow) place Systran, Reverso and Comprendium in a completely different order from the most frequent.

In terms of each system's ability to translate the different text types, all systems (including the human) obtained the highest fluency score for user manuals. This illustrates how well this text type is suited to translation by MT, and with the appropriate terminology in system dictionaries, a minimal amount of revision is required. For all but one system (FreeTranslation) the lowest fluency score was obtained for the FAQs. This is not surprising, bearing in mind the fact that the language used in these texts is often more

colloquial than in the other text types and in some instances is more akin to written chat, which is probably why it is more difficult for MT systems to handle. Yet the fact that adequacy scores were high for the FAQs meant that the lowest correlation between fluency and adequacy was found for this text type. This can be seen in Table 4.5. When examining scores for adequacy, no pattern was found in terms of the ability of systems to translate particular text types better than others. In fact, each MT system obtained its highest score for a different text type.

Having identified a high correlation between fluency and adequacy scores at system level (shown in 4.4.5.1), Pearson's correlation co-efficient was subsequently computed for the two attributes for each text type. Findings showed that with only around 1,200 words (3 documents of approximately 400 words of each text type) a high correlation was maintained between fluency and fidelity scores, adding weight to the theory that the evaluation of either fluency or adequacy is sufficient. Values are shown in Table 4.5.

Text type	Value of *r*
User manuals	0.9692
FAQs	0.9146
Press releases	0.9603
Technical reports	0.9846

Table 4.5: Correlation between fluency and adequacy scores by text type and system

4.4.6 Human evaluations: conclusions

The human evaluations described above have provided three scores for both fluency and adequacy for every segment in our sample of machine and human translations, generated from twelve source texts of approximately 400 words. This has provided essential corpus annotation and statistics for our research into the potential of a new automated scoring method for MT evaluation. Three significant conclusions can be drawn from the human evaluations:

1. Although based on a small sample, findings presented in 4.4.5.2 show that the evaluation of one text type (such as the news texts from the DARPA 1994 corpus)

cannot reliably indicate system performance for any text type. It is clear, therefore, that in a black box declarative evaluation, texts must be selected to reflect the user's intended use of MT.

2. The high correlation between fluency and adequacy scores, even with a smaller sample (ie. at the level of text type), is encouraging for this work: findings indicate that developing an automated evaluation method based on the analysis of the target text alone (like the human fluency evaluation conducted here) may well enable us to automatically predict adequacy scores too.

3. Results from the human evaluations show that there is no perceptible difference in system performance in some cases. In particular, Systran and Reverso produce very similar scores at system level: Systran outperforms Reverso for fluency by 0.006 and Reverso outperforms Systran for adequacy by 0.053. It his highly unlikely, therefore, that such minute differences will be detectable using an automated evaluation method with a sample of this size. In such instances, our aim will be to generate very close or equal automated scores.

4.5 Annotation and classification of errors

The second phase of corpus annotation involved the manual analysis of errors in the sample of MT output. An error classification scheme was required to enable the labelling of error types that typically occur in machine translated texts. Statistics from the annotations would then provide information on the frequency of error types produced by different MT systems, enabling us to compare error type frequency with human judgments of fluency and adequacy to guide the selection of errors for automated detection.

Existing translation error classification schemes were first explored to establish whether any of these might be appropriate or adaptable for this work. These are discussed in 4.5.1. As none were deemed suitable, a new fluency error classification scheme specifically for this research was developed. The development of the scheme is described in 4.5.2, along with practical issues and problems met during the manual annotation of errors. Examples of annotations are provided in 4.5.3. Frequencies of different error types are presented in 4.5.4

and findings from the error analysis are discussed in 4.5.5. Correlations between error frequencies and human quality judgements are investigated in 4.5.6 and conclusions are presented in 4.5.7.

4.5.1 Existing translation error classification schemes

Existing translation error classification schemes were first investigated to establish whether any of these might be suitable or adaptable for this work, bearing in mind that our analysis was to be based on the MT output alone, without access to the source text. It was found that most translation error classification schemes had been designed for a particular purpose. For example, the SAE J2450 Quality Metric, developed by the Society of Automotive Engineers (SAE J2450, 2001), the Framework for Standard Error Marking, devised by the American Translators Association (ATA, 2002) and BlackJack, ITR's commercially available translation evaluation method (ITR, 2006), were all designed for the evaluation of human translations. Many categories in these error classification schemes, shown in Appendix 5 are inappropriate for the annotation of errors in MT and they were found to be insufficiently fine-grained for the purpose of this work.

Very few error classification schemes designed specifically for MT were found, as fine-grained error analysis is not practical for the manual evaluation of machine translations. Such a method would be far more time-consuming than, for instance, the evaluation of fluency or fidelity at segment level. Correa's typology of errors commonly found in automatic translation (Correa, 2003) was unsuitable for our needs, largely because it was designed for glass-box evaluations during system development. A more useful starting point for this work was provided by: Flanagan's Error Classification for MT Evaluation (Flanagan, 1994), designed to enable end-users to compare translations by competing systems; Loffler-Laurian's typology of errors for MT, based on linguistic problems for post-editors (Loffler-Laurian, 1996); Wagner's error categories for end-users (Wagner, 1998) and classifications by Chaumier and Green in (Van Slype, 1979). However, all of these relied on access to the source text, which inevitably influenced the labelling of categories. These classification schemes are shown in Appendix 6.

4.5.2 Development of a new fluency error classification scheme

A data-driven approach was used to develop a fluency error classification scheme during the author's manual annotation of the English MT output from the four systems for which human scores had been obtained (Elliott et al., 2004b). This amounted to 48 machine translations and approximately 20,000 words in total. Provisional labels were initially assigned, according to items that a post-editor would need to amend if he/she were revising the texts to a publishable quality. Although no reference was made to the source text during this process, knowledge of the source language was necessary as, for instance, the scheme required untranslated words to be annotated with parts-of-speech. Furthermore, familiarity with NEs and acronyms (eg. names of software applications) in the texts was also required, to better represent the end-user of the translated texts and to code the terms appropriately.

Errors were annotated using the Systemic Coder (O'Donnell, 2005)[10], a tool that supports hierarchical linguistic coding schemes or ontologies and enables statistical analyses. Error types were sub-divided into parts-of-speech, as this would provide more fine-grained error groupings for analysis and would enable us to make more informed decisions when selecting errors for automated detection. As the Coder supports the insertion of new nodes into the hierarchy at any time, this facilitated the progressive data-driven refinement of the coding scheme. For example, after annotating around 1,000 words, a decision was taken to sub-divide 'inappropriate' items into 'meaning clear', 'meaning unclear' and 'outrageous' (words with an extremely low probability of appearing in a particular text type and subject area). This refinement would enable us to make better comparisons between MT systems, and isolate those errors that have a greater effect on intelligibility. The fluency error classification scheme is shown in Figure 4.16 (Elliott et al., 2004b).

The organisation of categories in the classification reflects the constraints of the coding tool to a certain extent. It was found during the annotation process that some coded items often involved two and, in rare instances, three error types. For example, a noun could be 'inappropriate', its position within the phrase could be incorrect and it could lack a required capital letter; a verb could be 'inappropriate' and the tense also incorrect. The scheme was,

[10] http://www.wagsoft.com/Coder/index.html

therefore, organized in such a way that the tool would allow all of these combinations of categories to be assigned to the same word or group of words where required.

Part of Speech	Compound noun	Noun string or Named entity	Inappropriate	Meaning clear/unclear/ outrageous
				Part meaning clear/unclear/ outrageous
			Untranslated	Untranslated
				Part untranslated
	Acronym	Incorrect		
	Pronoun	Inappropriate	Incorrect anaphor/other	
		Untranslated		
		Unnecessary		
		Omitted	Direct object pronoun	
			Relative pronoun	
			Other	
	Noun Adjective Adverb Conjunction	Inappropriate	Meaning clear/unclear/outrageous	
		Untranslated		
		Unnecessary		
		Omitted		
	Preposition	Inappropriate	With noun/verb/adjective	
		Untranslated		
		Unnecessary		
		Omitted		
	Determiner	Inappropriate		
		Untranslated		
		Unnecessary	Definite article	
			Indefinite article	
			Other	
		Omitted	Definite article	
			Indefinite article	
			Other	
	Verb	Inappropriate	Meaning clear/unclear	
			Outrageous	
			Multiword verb structure	
		Untranslated		
		Unnecessary		
		Omitted		
Tense or conjugation	Tense or mood			
	Conjugation			
Incorrect position	Acronym / pronoun / common noun / adjective / adverb / conjunction / preposition / determiner / verb / negator / noun string appendage			
	Compound noun sequence	Noun string or Named entity	Word order	
			Arrangement	
Other	Part of speech incorrect Inelegant or inappropriate style Incomprehensible expression Spelling error Incorrect negation Ordinal number untranslated Qualifier unnecessary			
	Number	Singular should be plural / plural should be singular		
	Case	Upper case required / lower case required		

Figure 4.16: Fluency error classification scheme

During the initial stages of text analysis, it became clear that, having set out to annotate fluency errors, adequacy errors were also detectable as contributors to disfluency, despite the absence of the source text. Words or phrases whose meaning was obviously incorrect in the given context were marked as 'meaning unclear' and can be regarded as both fluency and adequacy errors. For this research, we can, therefore, define each annotated error as a unit of language whose usage does not seem natural in the context in which it appears.

4.5.3 Examples of error annotation

The classification scheme enables any word or group of words to be annotated with up to four main error categories: part-of-speech, verb tense or conjugation, incorrect position and 'other'. Sub-categories are then selected, moving from left to right, until the final node is reached. Table 4.6 provides definitions of some of the categories with examples.

Error Type	Definition / Example
PoS: Verb Inappropriate Meaning clear	A verb is inappropriate in context but the meaning is understandable: …if the open file was already registered in this format … (saved)
PoS: Verb Inappropriate Meaning unclear	A verb is inappropriate in context and the meaning is not understandable The main window behaves the menu bar... (comprises)
PoS: Noun Inappropriate Outrageous	Intuitively, the word has a very low probability of appearing in this text type or domain How to define hourly beaches of use (time slots)
PoS: Verb Inappropriate Multiwd vb structure	A verb phrase comprising multiple words (in addition to prepositions) is incorrect The bans are more priority than authorizations (take priority over)
Incorrect position: Comp noun sequence NS: Word order	The constituent words of a noun string are ordered incorrectly access Internet is refused (internet access)
Incorrect position: Comp noun sequence Noun string Arrangement	The words of a noun string are ordered incorrectly and additional words are inserted – often prepositions and determiners - common in literal translations from French into English A window of definition of the filter opens (filter definition window)
Incorrect position Noun string appendage	Two noun strings are 'combined' in the SL so that when translated into English, the word order is incorrect. In this example, *tabs of options* would be marked as an incorrect NS arrangement, and the words in bold would be annotated as an incorrect noun string appendage position. the tabs of options **and regulations** (options and regulations tabs)
Incorrect position Adverb	Incorrect or unnatural positioning of an adverb a very great number of Francophone users avoids systematically using the characters highlighted (systematically avoids)
Determiner Omitted Definite article	Definite article omitted in cases where it is required in English. Three asterisks are inserted to mark the omission of an item. Among *** most frequent, … (the)

Table 4.6: Error types and definitions

4.5.4 Frequency of error types

Statistics were automatically generated from the annotation of errors in the 48 machine translations. Error frequencies for the 61 main categories are shown by system in Table 4.7.

Error type	Sys	Rev	Com	Free	Tot.
Nn inapp	78	80	81	105	344
Nn untrans	12	4	10	42	68
Nn unnec	0	2	2	5	9
Nn omitted	0	0	0	1	1
Pron inapp	14	36	23	31	104
Pron untrans	0	0	0	3	3
Pron unnec	9	6	11	16	42
Pron omitted	7	14	12	15	48
Acronym inc	7	11	9	8	35
NS inapp	52	42	71	85	250
NS untrans	0	0	0	0	0
NS inapp	8	11	12	23	54
NE untrans	5	4	11	5	25
Adj inapp	48	37	60	42	187
Adj untrans	2	1	3	12	18
Adj unnec	1	0	0	2	3
Adj omitted	0	0	0	0	0
Vb inapp	112	121	134	137	504
Vb untrans	8	1	0	6	15
Vb unnec	2	2	6	6	16
Vb omitted	4	3	4	11	22
Vb tense/mood	76	56	103	90	325
Vb conjugn	5	8	1	3	17
Prep inapp	73	77	84	89	323
Prep untrans	0	0	3	9	12
Prep unnec	25	26	49	41	141
Prep omitted	4	5	5	14	28
Det inapp	16	17	17	31	81
Det unnec	105	102	137	121	465
Det untrans	0	1	2	18	21
Det omitted	26	33	24	29	112

Error type	Sys	Rev	Com	Free	Tot.
Adv inapp	18	21	18	23	80
Adv untrans	0	0	0	4	4
Adv unnec	5	6	4	9	24
Adv omitted	0	0	0	0	0
Conj inapp	9	11	16	20	56
Conj untrans	0	0	0	0	0
Conj unnec	5	2	4	7	18
Conj omitted	2	3	6	2	13
Posn: Nn	5	4	3	8	20
Posn: Pn	2	2	3	9	16
Posn: Acron	0	0	0	0	0
Posn: Adj	17	15	24	42	98
Posn: Vb	6	8	6	16	36
Posn: Prep	2	0	1	2	5
Posn: Det	1	0	0	2	3
Posn: Adv	26	28	26	32	112
Posn: Conj	0	0	0	1	1
Posn: NS app	10	7	9	10	36
Posn: Neg	0	0	0	1	1
NS sequence	99	109	123	118	449
NE sequence	31	36	28	30	125
PoS wrong	11	6	8	32	57
Inelegant	22	24	14	22	82
Exp incomp	6	4	9	24	43
Spelling	2	2	0	1	5
Number inc	4	7	5	16	32
Case inc	24	22	24	29	99
Negation	0	0	1	4	5
Ord number	0	0	1	1	2
Qualif unnec	0	0	1	1	2
TOTAL	**1006**	**1017**	**1208**	**1466**	**4697**

Table 4.7: Error frequencies by system

4.5.5 Findings from error analysis

From the error frequencies shown in Table 4.7 a number of preliminary observations can be made. The mean number of errors per text (ie. around 400 words) across all systems was 97.85, and each segment (a sentence or heading) contained an average of 3.59 errors. The mean number of errors by system and text type is shown in Table 4.8.

Text type		Systran	Reverso	Comprendium	FreeTrans
All texts	Errors per text	83.83	84.75	100.66	122.16
	Errors per segment	3.07	3.11	3.69	4.48
User manuals	Errors per text	108.67	100.00	122.67	136.00
	Errors per segment	3.39	3.12	3.83	4.25
FAQs	Errors per text	85.33	84.33	97.00	125.67
	Errors per segment	3.01	2.97	3.42	4.43
Press	Errors per text	68.67	75.67	87.00	109.33
	Errors per segment	2.67	2.95	3.39	4.26
Tech reports	Errors per text	75.67	80.00	91.67	115.67
	Errors per segment	3.24	3.43	3.93	4.96

Table 4.8: Mean number of errors per text type

Findings show that for each system, more errors were found in the user manuals than in any other text type. There are two main reasons for this:

1. The user manuals contained a greater number of compound nouns, many of which were annotated as incorrect noun string sequences, the word order or arrangement of the noun string being incorrect. See Table 4.6 for error definitions and examples.
2. A large number of verbs were annotated as incorrect, although the meaning was still clear, because the word choice was inappropriate for the text type [eg. “safeguard”, “record” or “register” (a file) instead of “save”; “leave” (an application) instead of “exit”; “erase” instead of “delete”].

The FAQs were annotated with the second highest number of errors, their content being very similar to that of user manuals, with the addition of questions. Technical reports

contained the third highest number of errors and, for all systems, the lowest frequency of error was found in the press releases.

Inappropriate words represented the most frequent group of errors in the corpus sample, corresponding to 42.2% of all annotated errors (see Table 4.9). For all systems, verbs were the most frequent part-of-speech in this category, comprising 25.4% of the total number of errors in this group. Inappropriate nouns and prepositions were either the second or third most frequent parts-of-speech for each system, representing 17.3% and 16.2% of this error type. A pattern was also observed among other parts-of-speech marked as inappropriate: noun strings and adjectives were the fourth or fifth most frequent for each system, followed by pronouns, determiners or adverbs. Named entities or conjunctions were always the least frequent inappropriate words in output from any system.

Despite the large number of words labelled as inappropriate, in 88% of cases, the meaning was marked as clear. The majority of these items were considered inappropriate due to incorrect word choices in the given context or to unnatural collocations, rather than incomprehensibility.

Error type	Systran	Reverso	Comprend	FreeTrans	Total
Inappropriate word	428	453	516	586	**1983**
Incorrect word position	199	209	223	271	**902**
Unnecessary word	152	146	214	208	**720**
Incorrect tense	76	56	103	90	**325**
Omitted word	43	58	51	72	**224**
Untranslated word	34	22	39	108	**203**
Other (8 categories)	74	73	62	131	**340**
Total	**1006**	**1017**	**1208**	**1466**	**4697**

Table 4.9: Error groups in order of frequency

Incorrect word positions represented the second most frequent group of errors, comprising 19.2% of all annotated errors in the sample. The majority of items in this group (64%) were incorrectly arranged compound nouns (whether noun strings or named entities). Incorrectly positioned adverbs and adjectives were also relatively frequent, representing 12% and 11% respectively of errors of this type.

Unnecessary words comprised 15.3% of all annotated errors, the most frequent among these being determiners and prepositions, representing 65% and 19% of errors of this kind. This was not surprising given the more frequent use of these parts-of-speech in French compared with English. Determiners, for instance, are rarely omitted before nouns in French, whereas this is often the case in English; the preposition 'de' (of) is used to denote possession in French, when in English we often use the apostrophe 's' in these cases.

Other categories of error were considerably less frequent; incorrect tenses comprised only 6.9% of all errors, and the majority of these were found in the user manuals and FAQs, where imperatives were incorrectly translated as infinitives or gerunds. Omitted words represented just below 5% and untranslated words comprised 4.3% of all errors. Eight miscellaneous categories (marked with an asterisk in Table 4.7), each representing fewer than 100 errors within the corpus, were grouped together, comprising 7.2% of all annotated errors.

4.5.6 Comparing error frequencies and human judgements

Raw error frequencies were compared with scores from the human evaluations to explore whether any correlation existed between the number of annotated errors and intuitive judgments of fluency and adequacy. Assuming that all error types in the classification scheme affect fluency, the total number of errors for each system were initially compared with human fluency scores (see Table 4.10). All error categories that were considered unlikely to have an affect on adequacy were then removed (such as 'inappropriate' items with a clear meaning, unnecessary items, inappropriate prepositions and determiners, omitted determiners, incorrect positions of words, spelling errors, case errors and incorrect verb tense/mood or conjugation, the majority of these being an incorrectly translated imperative). The remaining classification of adequacy errors was then compared with the adequacy scores from the human evaluations, as shown in Table 4.10 (Elliott et al., 2004b).

System	Fluency		Adequacy	
	Mean human score and rank	No. of errors and rank	Mean human score and rank	No. of errors and rank
Systran	3.519 (1)	1006 (1)	4.136 (2)	121 (2)
Reverso	3.466 (2)	1017 (2)	4.142 (1)	120 (1)
Comprendium	3.221 (3)	1208 (3)	4.013 (3)	154 (3)
FreeTranslation	2.827 (4)	1466 (4)	3.644 (4)	279 (4)

Table 4.10: Comparison of fluency and adequacy scores with raw error frequencies

Findings showed that human scores for fluency and adequacy and the number of annotated errors rank all four systems in the same order. Pearson's correlation coefficient was calculated for the human scores and the number of errors per system for each attribute. A very strong negative correlation was found between the values: for fluency the value of r = -0.998 and for adequacy r = -0.981. Of course, only four pairs of variables are taken into consideration here. Nevertheless, findings indicate that adequacy as well as fluency judgements can be captured based on the frequency of errors annotated without reference to the source text. This preliminary observation is encouraging for the development of an automated approach to error detection which is based only on the MT output.

The number of fluency errors and the sub-set of adequacy errors were then compared with human scores for fluency and adequacy according to text type. No significant correlation was found (Elliott et al., 2004b). In fact, human scores for fluency were highest for the user manuals for all systems (see Table 4.4), yet these texts contained the largest number of annotated errors. As discussed in 4.5, the two main reasons for the large number of errors annotated in the user manuals were (i) the high frequency of compound nouns which, in many cases, were coded with two error types (eg. inappropriate translations and word order) and (ii) the high number of inappropriately translated verbs, which although understandable in the majority of cases, were not correct in the context of software applications. Furthermore, user manuals were annotated with the largest number of untranslated words, yet many of these could be understood by evaluators with no knowledge of French, having little or no effect on human judgements of adequacy. These intuitions were supported by a further experiment in which 58% of all untranslated words in the sample were correctly guessed in context by three people with no knowledge of French.

In fact, when all untranslated words from the sample were presented in the form of a list, 44% were correctly guessed out of context. It is clear, therefore, that some error types should be weighted to correlate with intuitive human judgements of translation quality.

4.5.7 Error annotation and classification: conclusions

An adaptable, hierarchical fluency error classification scheme has been developed for French-English MT output, which also enables the detection of adequacy errors, without access to the source text. A total of 4,697 errors were annotated in approximately 20,000 words of MT output, equal to approximately 97.85 errors per text (of around 400 words) and 3.59 errors per segment. The annotation of output from four systems has shown that raw error frequency by system correlates highly with human judgments for fluency, and that a sub-set of error categories correlates with human judgments for adequacy. Without error weighting, there is no correlation between error frequency and human fluency or adequacy judgements at the level of text type within our sample.

4.6 Corpus compilation: conclusions

A new, publicly-available, parallel corpus has been developed specifically for French-English MT evaluation. This will be updated with texts in other languages as they are required. Due to constraints in terms of time and costs, a detailed rationale for corpus design was first developed, based on responses to an MT user survey and experiments with texts from the DARPA corpus to determine sample size. The forty source texts reflect commercial use of MT, and each has four machine translations, one human translation and one reference translation.

A sample of sixty translations from the corpus (twelve texts translated by four MT systems and one human) has been annotated at the segment level with human scores for fluency and adequacy from three different judges, obtained from evaluations based on the DARPA 1994 series. Segmented texts and their associated scores are available as .txt files, and the collated scores are also available in an Excel spreadsheet. The forty eight machine translations (approximately 20,000 words) have also been annotated with error categories based on a new hierarchical error classification scheme designed specifically for MT output, which is used without access to the source text or to any human translation. These

annotations are viewable in separate files via the freely-downloadable Systemic Coder[11] which also enables statistical analyses of the data.

Annotations representing fluency and adequacy judgements are essential to enable us to investigate the validity of scores from a new automated evaluation method, and statistics generated from the manual error annotations have provided valuable information for the investigation of the potential of an automated approach to error detection, which relies on the MT output alone.

[11] http://www.wagsoft.com/Coder/index.html

Chapter 5

Automated error detection

Having explored existing automated MT evaluation methods, their successes and weaknesses in Chapter 3, this chapter begins in 5.1 with an introduction to the new automated approach, including the advantages of automated error detection for MT evaluation, the aims of the new method and the selection of error types for initial automatic detection. Methods for the detection of five selected error types are then described in 5.2, along with results and correlations with human scores. In Section 5.3, scores obtained from the detection of the five error types are combined to investigate whether closer correlations are obtainable with human judgements when more than one error type is taken into account. Conclusions on automatic error detection are presented in 5.4.

5.1 The development of a new approach

A reliable evaluation method based on the automatic identification of errors, which requires access only to the MT output, would be of use to many MT stakeholders. The advantages of the proposed method are outlined in 5.1.1, and the specific aims during the development of the new approach are presented in 5.1.2. The selection of error categories for initial automatic detection is discussed in 5.1.3.

5.1.1 Advantages of the proposed method

Unlike existing approaches to automated evaluation described in 3.5, the proposed method would generate two and, in some cases, three possible outputs: annotated errors in the machine translations, an automated score for a given sample of MT output (which should ultimately be able to predict human judgements of translation quality) and, where possible, suggested error corrections. The chosen approach has a number of advantages:

- As the proposed method relies solely on the MT output, this eliminates the need for human translations, which are expensive to produce and, unlike most existing automated methods (excluding those described in 3.5.4 and 3.5.5), this enables users to evaluate any text sample without relying on pre-translated source texts.

- The method would be of use to post-editors who need to quickly identify particular errors for revision.
- It would enable the user to quickly view errors such as untranslated or 'inappropriate' content words for dictionary update.
- Unlike existing automated methods, such as BLEU (Papineni et al., 2001), it would help users or purchasers to compare the performance of competing systems based on error information as well as automated scores. BLEU provides nothing but a score for a given document; it provides no information on translation problems.
- The method would help developers to pinpoint areas for improvement.
- Such an approach should be extendible to other language pairs in which the target language is English. Results from the testing of algorithms on other language pairs are reported in Chapter 6.
- As the method relies on a suite of independent algorithms, rules for detecting different error types could be switched on or off at any time, allowing the user to focus on particular error types.
- Algorithms for the detection of different errors could be added or amended at any time. This modular approach facilitates experiments in the use of weighting for different error types to enable the constant improvement of correlations between automated scores and human judgements.
- Unlike BLEU (Papineni et al., 2001), the new approach should ultimately be able to take the gravity of each error type into account by introducing weighted scores, rather than assuming that all tokens in a text are of equal importance.

5.1.2 Aims of the new approach

In Section 3.5.7.1 a list of desiderata for automated MT evaluation was presented, based on observations from an investigation of existing automated methods. With these points in mind, aims during the development of the proposed approach are as follows:

(a) Scoring

a. The ultimate aim is to produce finely differentiated scores, rather than just ranks, based on output from competing systems of varying levels of quality;

b. Automated scores should be predictive of human judgements and correlations with human scores for both fluency and adequacy are to be explored;

(b) Accuracy

a. The aim is to design algorithms that detect as many errors as possible within each error group;
b. Attempts will be made to prevent the incorrect flagging of correct items.

(c) Extendibility

a. It is intended that the method will be extendible to other language pairs in which the target language is English. This involves the testing of algorithms on MT output from other source languages to determine their portability and adaptation requirements;
b. Error detection should be adaptable to personal needs, so the aim is to incorporate a feature to enable the user to enter new terms (eg. named entities) that should not be flagged as errors.
c. It is intended that some algorithms, which detect particular error types, will enable automatic error correction to help post-editors. According to (Allen, 2003) "a post-editor is likely to have to fix the same errors again and again in daily post-editing tasks. (...) For post-editing, it would be desirable to have a processing engine that could automatically fix up the highly frequent, repetitive errors in raw MT output before such texts are even given to a human post-editor." No fully automated method has been developed to date. The APE system (Povlsen and Bech, 2002) *automatically* learns from previously post-edited documents and enables the *semi-automatic* correction of the most common repetitive errors. In contrast, the algorithms explored here are developed based on manual text analyses, and tested to ensure that very few exceptions to the rules occur, enabling the programs to be fully automatic.

5.1.3 Rationale for the selection of errors for initial automatic detection

Having created a classification of errors, as described in 4.5.2, the next goal was to select error categories for which automatic detection methods would initially be explored. In terms of selecting error types, the following points were considered:

a) the frequency of errors in each category;
b) the error types found in segments with the lowest human scores; ie. assumed to represent the most serious errors;
c) anticipated computational tractability.

Strong correlations between the frequencies of manually annotated error types and human scores were not considered as criteria for the selection of errors for automatic detection, as it would be highly unlikely that all errors in a given category would be automatically detected. A major reason for this is that many errors are surrounded by other incorrect items, making them very difficult or impossible to detect when relying on context. Furthermore, we cannot be certain that false positives will not occur during the detection of errors in some categories.

With regard to (a) above, the error types listed in Table 4.7 were grouped together and frequencies were calculated: these are shown in Table 5.1, in which 'inappropriate' words represent the largest group.

Error type	Total	% of total number of errors
Inappropriate word	**1983**	**42.22**
Incorrect word position	**902**	**19.20**
Unnecessary word	**720**	**15.33**
Incorrect tense	**325**	**6.92**
Omitted word	**224**	**4.77**
Untranslated word/acronym	**203**	**4.32**
Other (8 categories)	**340**	**7.24**
Total	**4697**	**100.00**

Table 5.1: Error types and frequencies

It is essential, therefore, that we begin by attempting to detect some of the errors of this type. In contrast, the eight categories listed under 'Other' (incorrect verb conjugations, incorrect parts-of-speech, inelegant style, incomprehensible expressions, spelling errors, incorrect number, incorrect case and negation) were not considered for detection, as each category comprises a comparatively small number of errors.

With regard to (b) above, segments given the lowest scores by human judges were analysed to determine the types of errors found within them. For fluency, 76 segments (0.58% of the sample) received scores of between 3 (the lowest possible score) and 5 out of a possible 15, from the three evaluators; for adequacy, 64 segments (0.49% of the sample) received scores of between 3 and 6. Table 5.2 shows the mean number of each error type per segment in the whole sample compared with the mean number found in the lowest scoring segments for fluency and adequacy.

Error types	Errors in whole sample (1308 segts)	Mean no per segment in whole sample	Errors in 76 segts scoring <6 for fluency	Mean no per segment and variation from whole sample	Errors in 64 segts scoring <7 for adequacy	Mean no per segment and variation from whole sample
Inappropriate word	1983	1.52	190	2.50 x1.64	124	1.94 x1.28
Incorrect position	902	0.69	81	1.07 x1.54	59	0.92 x1.34
Unnecessary word	720	0.55	75	0.99 x1.79	30	0.47 x0.85
Incorrect tense	325	0.22	31	0.41 x1.64	11	0.17 x0.69
Omitted word	224	0.17	26	0.34 x2.00	7	0.11 x0.64
Untranslated word/acronym	203	0.15	27	0.35 x2.33	22	0.344 x2.29

Table 5.2: Error types in segments with lowest human scores

In the lowest scoring segments for fluency, the mean number of errors in every category increased to between 1.54 and 2.33 times the mean number in the entire sample, as shown in Column 5 of Table 5.2. However, in the lowest scoring segments for adequacy, the frequency of only three of the error types increased from the mean: inappropriate words,

incorrect word position and untranslated words. These are highlighted in the right-hand column of Table 5.2. This shows that unnecessary words (84% of which were determiners and prepositions), incorrect tenses (mostly inappropriate use of the present tense or infinitives) and omitted words (84% of which were pronouns, prepositions and determiners) were not major contributors to low fidelity scores. In contrast, the mean number of untranslated words in the lowest scoring segments for both fluency and adequacy more than doubled when compared to the mean number in the whole sample.

With this in mind, a decision was taken to focus initially on the semantic (inappropriate and untranslated words) rather than the syntactic errors: semantic errors were considered to be more serious, as they were found to be significantly more frequent in segments given lower scores by humans, and this choice would enable the development of algorithms without the need for parsers or part-of-speech taggers, making the methods more resource-light and, therefore, more easily replicable. For this reason, incorrect position errors, although significantly more frequent in the lowest scoring segments, would be the focus of future work.

Five groups of errors were selected, based on their expected computational tractability. These were categorised according to the different approaches anticipated for their detection:

1) Inappropriate words marked as "outrageous";
2) Inappropriate words marked as "meaning clear" or "meaning unclear";
3) Inappropriate prepositions;
4) Inappropriate pronouns;
5) Untranslated words and acronyms.

This selection includes all inappropriate content words, 76% of inappropriate function words and all untranslated items, and involves the investigation of 44% of errors in the whole sample.

5.2 Error detection algorithms

For each error type, the aim was to devise an automated detection method, based on human processing rationales developed during manual text analyses. Each algorithm was evaluated by comparing the number of automatically detected errors in each category with manual annotations. The number of errors detected for each error type was then compared with human judgements to investigate correlations between the two sets of data.

For each error type, the development of the detection method is first described and results are presented to show the accuracy of each algorithm, based on comparisons between automatically and manually annotated errors. Correlations between automated and human scores are then investigated. The automatic detection of 'outrageous' words is described in 5.2.1, other 'inappropriate' words in 5.2.2, 'inappropriate' prepositions in 5.2.3, pronouns in 5.2.4 and untranslated words and acronyms in 5.2.5.

5.2.1 Automatic detection of outrageous words

Among the 1419 content words manually annotated as 'inappropriate', 1,180 were labelled as 'meaning clear', 187 as 'meaning unclear' and 52 as 'outrageous'. Annotated items were further categorised according to their part-of-speech, as shown in Table 5.3, which may prove useful for future error weighting.

Part of speech	Systran				Reverso				Comprendium				FreeTranslation				TOTAL
	Meaning clear	Meaning unclear	Outrageous	TOTAL	Meaning clear	Meaning unclear	Outrageous	TOTAL	Meaning clear	Meaning unclear	Outrageous	TOTAL	Meaning clear	Meaning unclear	Outrageous	TOTAL	
Verb	107	5	0	112	110	8	3	121	120	12	2	134	128	7	2	137	**504**
Noun	59	13	6	78	60	16	4	80	58	16	7	81	77	26	2	105	**344**
NS	33	16	3	52	32	9	1	42	53	14	4	71	68	12	5	85	**250**
Adj	40	5	3	48	32	1	4	37	48	9	3	60	33	8	1	42	**187**
Adv	17	1	0	18	20	1	0	21	17	1	0	18	21	2	0	23	**80**
NE	8	0	0	8	9	2	0	11	10	1	1	12	20	2	1	23	**54**
TOTAL	264	40	12	316	263	37	12	312	306	53	17	376	347	57	11	415	**1419**

Table 5.3: Number of annotated inappropriate words by system

Some of the more serious semantic errors among those shown in Table 5.3 - the outrageous words - were targeted first. These words, by definition, are highly unusual in a given text type and domain; they also give no clue to the intended meaning in the translation. In most cases, outrageous words are mistranslations which normally occur when a source language word has two or more valid translations in the target language, with very different meanings. Rare exceptions are untranslated words, legitimate in both the source and target language, but deemed highly inappropriate in the target text. One example of this is the word 'moult' shown in Table 5.4.

The majority of outrageous words (43 out of 52) had been annotated in two of the four text types from the sample: the software user manuals and FAQs on software topics, and were found to apply to verbs, nouns, noun strings, named entities and adjectives. Examples are shown in Table 5.4. To the author's knowledge, no method has yet been developed for the automatic detection of this type of error in any kind of translation.

French word	MT output	Required translation
plage	beach	time slot
onglet	mitre	tab
standard	switchboard	standard
buse	buzzard	pipe/nozzle
hébergeur	shelterer	(web) host
interdit	speechless	prohibited
moult	moult (untranslated)	many
collant	tights	pasting
favoris	sideburns	favourites (on web)

Table 5.4: Examples of annotated outrageous words

Rather than hardwiring these words for automatic error detection, which would prevent the capture of additional outrageous words in unseen data, experiments were conducted to determine whether these tokens might be detectable by automatically comparing the MT output against a lexicon generated from a corpus of original English texts in the same genre and domain; words found in the MT output but not in the original English texts may provide a list of potential outrageous words to be flagged as errors. However, untranslated words would also be flagged using this method, so an English lexicon would also be required to filter out these tokens (see 5.2.1.2).

5.2.1.1 Compiling corpora for the detection of outrageous words

First, a three million word English corpus of software user manuals and FAQs on software topics was compiled from the Web with the aid of the BootCaT Toolkit (Baroni and Bernadini, 2004). This was used to generate a 'software lexicon', which was then enhanced with additional words and acronyms/abbreviations from FOLDOC (Free On-Line Dictionary of Computing[1]) to provide a corpus of assumed acceptable words for the computer software domain.

A lexicon of English was then compiled to be used for the filtering out of untranslated words. The complete, unlemmatised frequency list from the written part of the British National Corpus[2] was found to be the most extensive list of words used in written (British) English, available in electronic form. The 'word' column, containing 921,074 entries, was first extracted to provide the basis for our lexicon. The text was then converted to lower case. In a first step to reduce the size of the file, and to ultimately speed up processing with the final algorithm, duplicate entries were deleted to create a unique word list. (Many tokens appeared more than once, labelled with different parts-of-speech.) All non-words[3] were also deleted.

In addition to "general" English words, the MT output was also expected to include domain-specific terms from the field of computing, so the word list was then enhanced with additional tokens from the software lexicon described above. At this stage it contained 323,248 entries: around one-third of its original size. The intention is to progressively add new tokens to the file as new words and acronyms are identified in future evaluations.

5.2.1.2 Automatic detection of outrageous words: method 1

The following algorithm (Elliott et al., 2005) was used to compare words in the MT output against the software lexicon, and to tag all potential outrageous words (ie. those which did not appear in the lexicon) in the translations.

---

[1] Available at http://foldoc.doc.ic.ac.uk/foldoc/index.html

[2] Freely downloadable from the ITRI website: http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html#raw

[3] For the purpose of this thesis, non-words refer to any character string containing one or more non-alphabetic characters (excluding punctuation). These include email addresses, URLs and file names.

(1) Convert copy of MT output file to lower case, in accordance with lexicon;
(2) Remove all 'non-words' from MT that do not require analysis (tokens containing non-alphabetic characters: URLs, email addresses, file names, dates, etc.) and replace hyphens with a space, so that constituent parts are identified separately in the lexicon, and are not incorrectly flagged;
(3) Tokenise MT output using Natural Language Toolkit (NLTK) Tokeniser (Loper and Bird, 2005);
(4) Create a unique word list of all remaining tokens;
(5) Compare unique word list against lexicon from user manual/FAQ corpus to produce list of tokens found in MT but NOT in lexicon;
(6) Compare new list against English lexicon from the BNC to filter out untranslated words;
(7) For each word in the final list, tag all occurrences of that word (regardless of case) in the *original* MT output file.

5.2.1.3 Method 1 results

The algorithm was applied to detect and annotate outrageous words in the 10,000 words of machine translated user manuals and FAQs. Automatically flagged tokens and manual annotations for all inappropriate content words were then compared. Results for all four MT systems (Reverso, Systran, Comprendium and FreeTranslation) are shown in Figure 5.1 (Elliott et al., 2005). Findings showed that all manually annotated outrageous words were automatically identified as such. Furthermore, 26% of inappropriate words labelled as 'meaning clear' and 46% of words labelled as 'meaning unclear' were also automatically flagged. However, the large number of incorrectly flagged items in the MT output, also shown in Figure 5.1 meant that this method had to be rejected.

In order to reduce or eliminate the number of false positives, a second detection method was formulated and tested: the decision was taken to build an updateable lexicon of outrageous tokens found in a very large sample of machine translated texts in the computer software domain, with the aim of increasing the ability to capture errors in unseen data. Words would be admissible to the lexicon only if all occurrences of that token in the sample were deemed highly inappropriate. This method is described in 5.2.1.4.

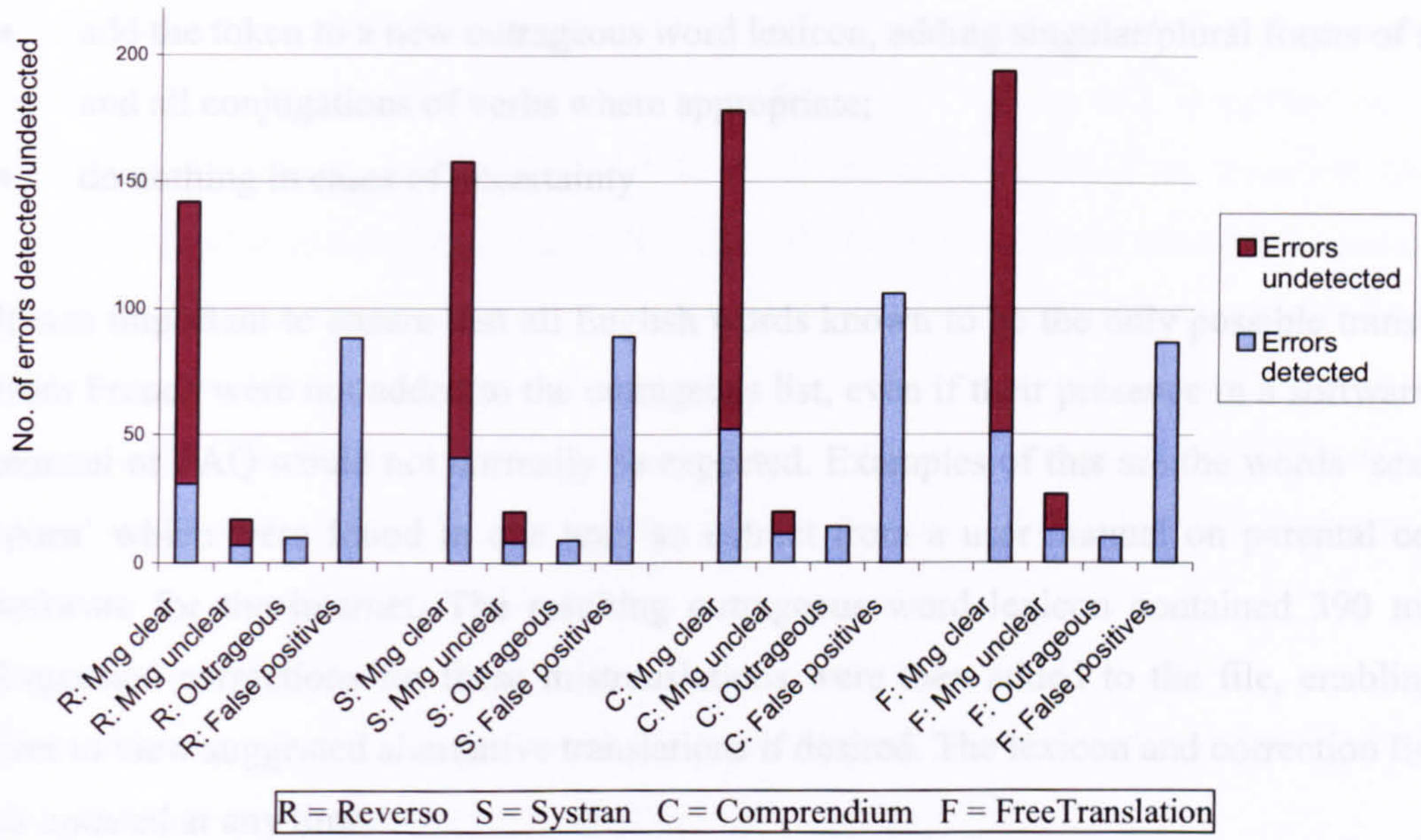


Figure 5.1: Results from automatic detection of outrageous words (Method 1)

5.2.1.4 Automatic detection of outrageous words: method 2

A 250,000 word corpus of previously unseen French user manuals and FAQs was compiled from the Web. This was then machine translated by three available commercial systems: Systran, Reverso and Idiomax. Free online systems were not used, as they require the user to paste limited amounts of text into a small box, which would be laborious and extremely time-consuming.

For each of the three samples of MT output, all tokens not requiring analysis (ie. those containing non-alphabetic characters) were automatically discarded. A unique word list in lower case was then generated from all three samples. This was then automatically compared against the software lexicon (see 5.2.1.1), returning a list of all tokens not found. To reduce the size of the file, all tokens not found in our English lexicon (ie. untranslated words) were automatically filtered out. These tokens were manually checked, and valid words, many of which were American spellings, were added to the English lexicon. The remaining list of potential outrageous words contained 2,998 tokens.

Each token in the list was manually examined and a decision made to:

- add the token to the software lexicon if deemed acceptable;

- add the token to a new outrageous word lexicon, adding singular/plural forms of nouns and all conjugations of verbs where appropriate;
- do nothing in cases of uncertainty

It was important to ensure that all English words known to be the only possible translation from French were not added to the outrageous list, even if their presence in a software user manual or FAQ would not normally be expected. Examples of this are the words 'sex' and 'porn' which were found in one text: an extract from a user manual on parental control software for the internet. The resulting outrageous word lexicon contained 390 tokens. Suggested corrections for these mistranslations were then added to the file, enabling the user to view suggested alternative translations if desired. The lexicon and correction list can be updated at any time.

5.2.1.5 Method 2 results

Our annotated 10,000 word sample of machine translated user manuals and FAQs was automatically compared against the outrageous word lexicon and all matches flagged. Automatically flagged tokens and manual annotations were again compared. Results by system are shown in Figure 5.2 (Elliott et al., 2005).

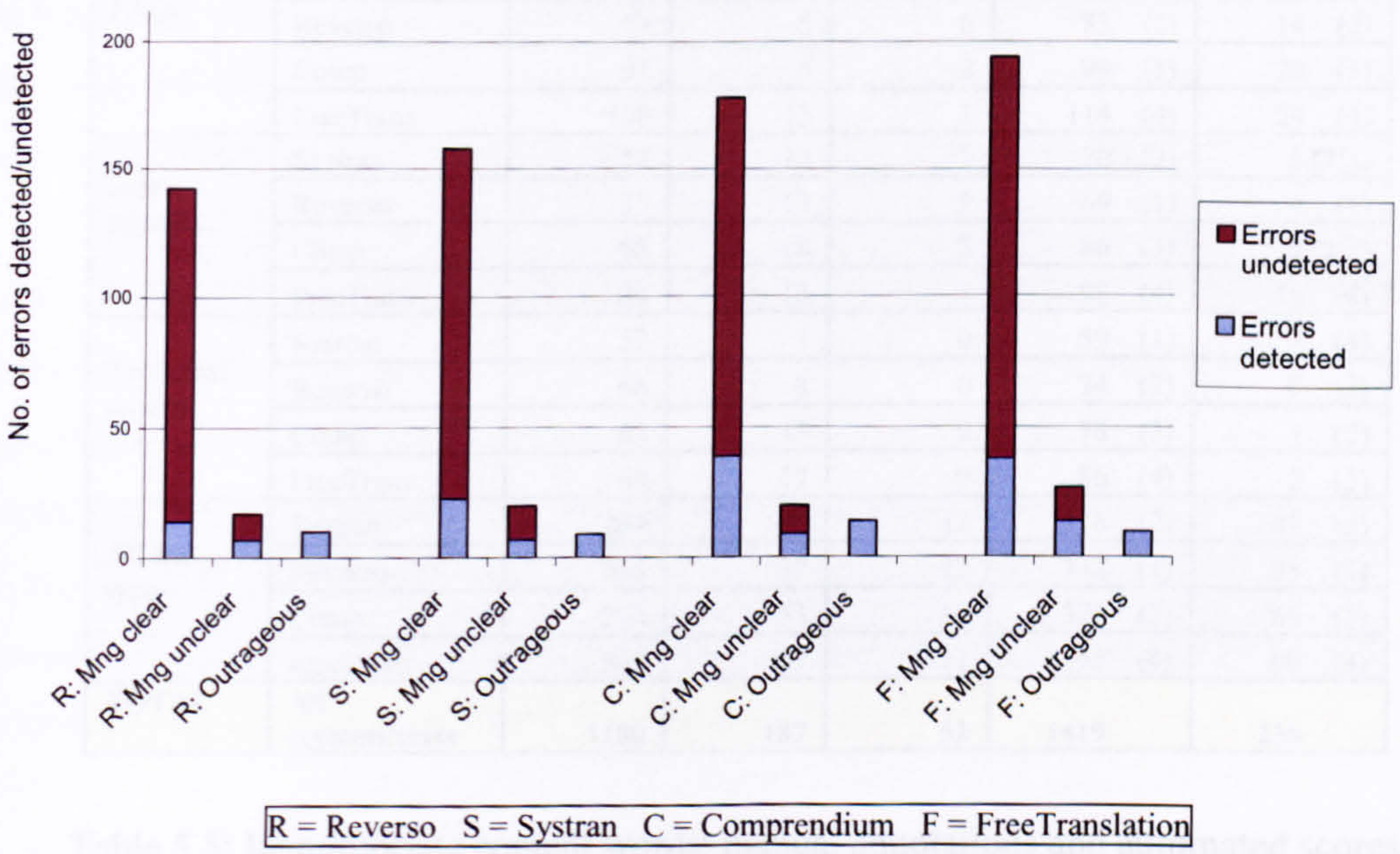


Figure 5.2: Results from automatic identification of outrageous words (Method 2)

Using this new method, findings showed that all 43 manually annotated outrageous words were automatically identified as such. In addition, 114 of the 672 inappropriate words labelled as 'meaning clear' (17%) and 37 of the 84 words labelled as 'meaning unclear' (44%) were also automatically flagged. Although these two figures are slightly lower than for Method 1, the detection of some errors in these categories remains valuable; it had been envisaged that all errors in these two groups would need to be automatically detected in a completely different way from the outrageous words, by taking context and collocation expectations into account. The main advantage of Method 2 over Method 1 is that no false positives were flagged in the sample data.

Method 2 was subsequently tested on the press releases and technical reports. Findings for all text types are summarised in Table 5.5.

Text type	System	Manual annotations: Meaning clear	Meaning unclear	Outrag-eous	Total no. of manual annotations and rank	Auto score (number of items detected) and rank
User manuals	Systran	86	14	5	105 (2)	27 (2)
	Reverso	68	12	4	84 (1)	17 (1)
	Comp	87	14	12	113 (3)	42 (4)
	FreeTrans	94	14	9	117 (4)	38 (3)
FAQs	Systran	72	6	4	82 (1)	12 (1)
	Reverso	74	5	6	85 (2)	14 (2)
	Comp	91	6	2	99 (3)	20 (3)
	FreeTrans	100	13	1	114 (4)	24 (4)
Press releases	Systran	54	13	3	70 (2)	5 (2/3)
	Reverso	55	12	2	69 (1)	4 (1)
	Comp	65	18	3	86 (3)	5 (2/3)
	FreeTrans	84	13	1	98 (4)	16 (4)
Technical reports	Systran	52	7	0	59 (1)	3 (4)
	Reverso	66	8	0	74 (2)	0 (1)
	Comp	63	15	0	78 (3)	1 (2)
	FreeTrans	69	17	0	86 (4)	2 (3)
All text types	Systran	264	40	12	316 (2)	47 (2)
	Reverso	263	37	12	312 (1)	35 (1)
	Comp	306	53	17	376 (3)	68 (3)
	FreeTrans	347	57	11	415 (4)	80 (4)
TOTAL	**All systems/texts**	**1180**	**187**	**52**	**1419**	**230**

Table 5.5: Inappropriate content words: manual annotations and automated scores

The press releases and technical reports contained only 9 words manually annotated as 'outrageous' and all were automatically detected along with a small number of other 'inappropriate' words: 17 out of 508 (3.3%) of inappropriate words marked 'meaning clear' and 10 out of 103 (9.7%) of words marked 'meaning unclear'. No words were incorrectly flagged.

Results also show that the number of errors detected reflects the rank ordering of systems in terms of the number of manually annotated errors when all text types are considered. However, this is less often the case when error counts for individual text types are taken into account. Coinciding ranks based on manual and automated annotations are highlighted in Table 5.5.

5.2.1.6 Correlations with human judgements

Automated scores, representing the number of 'outrageous' words detected using Method 2, were compared with human judgements for fluency and adequacy (shown previously in Tables 4.3 and 4.4). Initial observations showed that the number of detected 'outrageous' words reflected the rank order of systems according to human adequacy judgements when all text types were taken into consideration. Automated scores could not, however, reflect ranks based on both adequacy *and* fluency judgements for the whole sample, as the human scores order the systems differently for each attribute, due to very close competition between the two top-scoring systems. No rank correlations were found at the level of individual text types.

Using Pearson's correlation coefficient, the strongest correlations were again found when comparing automated scores for all text types with human judgements. Correlations for individual text types were much weaker in all cases and extremely low for the technical reports in which no 'outrageous' words had been manually annotated and very few other 'inappropriate' words were automatically detected. Across the whole sample, automated scores were found to correlate more strongly with human judgements of fluency rather than adequacy. Correlation coefficients are shown in Table 5.6.

Texts	Value of *r* Automated scores: fluency	Value of *r* Automated scores: adequacy
User manuals	-0.7119	-0.5105
FAQs	-0.7350	-0.8412
Press releases	-0.8281	-0.8379
Tech reports	-0.0038	0.1113
All text types	**-0.9125**	**-0.8774**

Table 5.6: Correlations: detected outrageous words and human judgements

5.2.1.7 Conclusions

An algorithm was developed which, with the aid of two purpose-built lexicons, and with access only to the MT output, was able to detect all words manually annotated as 'outrageous' in the sample. In addition, the algorithm detected a number of other words marked as 'inappropriate' and no items were incorrectly flagged. In total, 230 of the 1419 manually annotated 'inappropriate' content words were detected using this method, equal to 16.2%. Furthermore, this approach enabled the algorithm to be extended to offer possible error corrections to the user.

The selected method was based on the detection of words whose appearance is highly unusual in a particular text type or domain; other 'inappropriate' words should be detectable by methods based on the more immediate context (ie. by carrying out analyses at the phrase or sentence level).

Correlations between human and automated scores were encouraging, but the detection of additional error types would be required to enable the investigation of stronger correlations with human scores.

5.2.2 Automatic detection of other inappropriate content words

After the detection of 'outrageous' words, 84% of the inappropriate content words remained undetected. Unlike 'outrageous' words, which can be automatically detected due to their highly unusual appearance in particular text types and domains, the remaining

errors in this category were defined as inappropriate for a number of different contextual reasons, such as those shown in Table 5.7.

Problem	Example
Unnatural collocation: eg. wrong choice of adjective or adverb; meaning usually clear	**principal** menu / the **good** procedure / **a big** number of read **attentively** / see **higher**
Verb incorrect due to contextual expectations; meaning usually clear	**record** the file (save) / **keep in memory** the document (save/store) **register** the file (save) / **leave** the program (quit/exit)
Incorrectly (often literally) translated noun string; meaning often clear	word of past (password) / mice click (mouse click) guide of use (user guide)
Error due to incorrect parse or lexical choice when SL word has more than one translation possibility: meaning often unclear	the **running** document (current) / **treatment** of texts (processing) **waiter** centre (server) / a **good-hearted** problem (minor) **a spilled** example (widespread) **small file** (file menu) – SL word 'menu' has been parsed as an adjective
Inelegant choice of word or expression	listed in the list / carefullest / the possessor of

Table 5.7: Examples of inappropriate content words

Having automatically identified 16.2% of inappropriate content words using the algorithm described in 5.2.1.4, 1189 of these errors remained undetected. Table 5.8 shows the number of remaining errors by system, 88% of which were labelled as 'meaning clear'.

System	Meaning clear	Meaning unclear	Total number of annotations
Systran	238	31	269
Reverso	249	28	277
Comp	266	42	308
FreeTrans	294	41	335
All systems	**1047**	**142**	**1189**

Table 5.8: Remaining inappropriate content words by system

5.2.2.1 Development of the detection method

The remaining inappropriate content words in the sample were subsequently analysed in context. From this analysis it became clear that many of the same, or similar, errors were occurring regularly in output from more than one and, in some cases, from all four MT systems. These include some of the examples shown in Table 5.7. With knowledge of the source language, the cause of error was clear in many cases, often attributable to an incorrect lexical choice in the given context when a SL token has more than one legitimate translation.

With this in mind, a decision was taken to compile a 'reverse' test suite (see 3.4.9), comprising a list of unacceptable word combinations found in a large sample of MT output from different systems, against which our smaller sample could be automatically compared and error matches flagged. As a starting point, a sample of 90,000 words from the corpus of machine translated software user manuals and FAQs (see 5.2.1.4) was manually analysed. This comprised translations of the same 30,000 words (approximately) of French source texts by each of three MT systems. Incorrect word combinations, including examples shown in Table 5.7, were selected from the MT output and added to an inappropriate words list. Each involved one item considered to be inappropriate. Checks were made to ensure that none of these errors were already contained in the 'outrageous' word list, to prevent any item from being detected twice.

In an attempt to make the method more robust, inappropriate or partly inappropriate noun strings and named entities were not included, as their correct translations could be added at any stage to an MT system's user dictionary; focusing on other kinds of incorrect word combinations, which are not so easy to rectify, should make results from this method more robust in the long term.

Furthermore, some other word combinations were not selected for the rules file if, for example, they involved a word that was considered appropriate in some contexts but inappropriate in others. For instance, the verb 'comprise' and its different conjugated forms was acceptable in some phrases but not in other contexts, these judgements being based on real world knowledge. In (1) below, the verb was considered to be inappropriate (a

preferable translation would be 'contain') whereas in (2) the verb was not deemed inappropriate.

(1) French texts **comprise** accents.

(2) A Postscript file **comprises** two parts.

Although this verb was frequent in the analysed sample of MT output, it was found to combine with a very large number of other words; adding the erroneous combinations to the rules file would not be efficient, as these collocations might never be found in other samples. Instead, words involving more frequent collocations were selected, such as 'register the file' or 'register the document'.

Some inappropriate content words were found in word combinations involving other types of error, such as untranslated words or incorrect word ordering. These were not added to the list as (a) each instance was expected to be infrequent and (b) some of the other adjacent errors could be remedied at any time during the system's development, making these items in the rules file redundant.

Many of the word combinations selected for the list were subsequently searched in the BNC to determine whether or not they are legitimate in the English language and, if so, in which domains and text types they tend to occur. Based on these findings, items that were expected to generate false positives, due to their legitimate occurrence in some contexts, were removed from the list. Having finalised the word combinations for inclusion, steps were taken to ensure that for items containing verbs, these were listed in all their forms and that for items involving nouns, these were included in both their singular and plural forms where appropriate. The list was then formatted as a file of rules and tags to be automatically inserted into the MT output at the appropriate points of error. The final list contained 1,490 entries. Suggested error corrections could also be added to each entry in the file if required, for automatic insertion into the MT output. This is an area for future work.

5.2.2.2 Results

The 20,000 word sample of MT output was automatically compared against the rules file and all errors found were tagged. Automatically detected tokens and manual annotations were compared. Findings by system are shown in Figure 5.3.

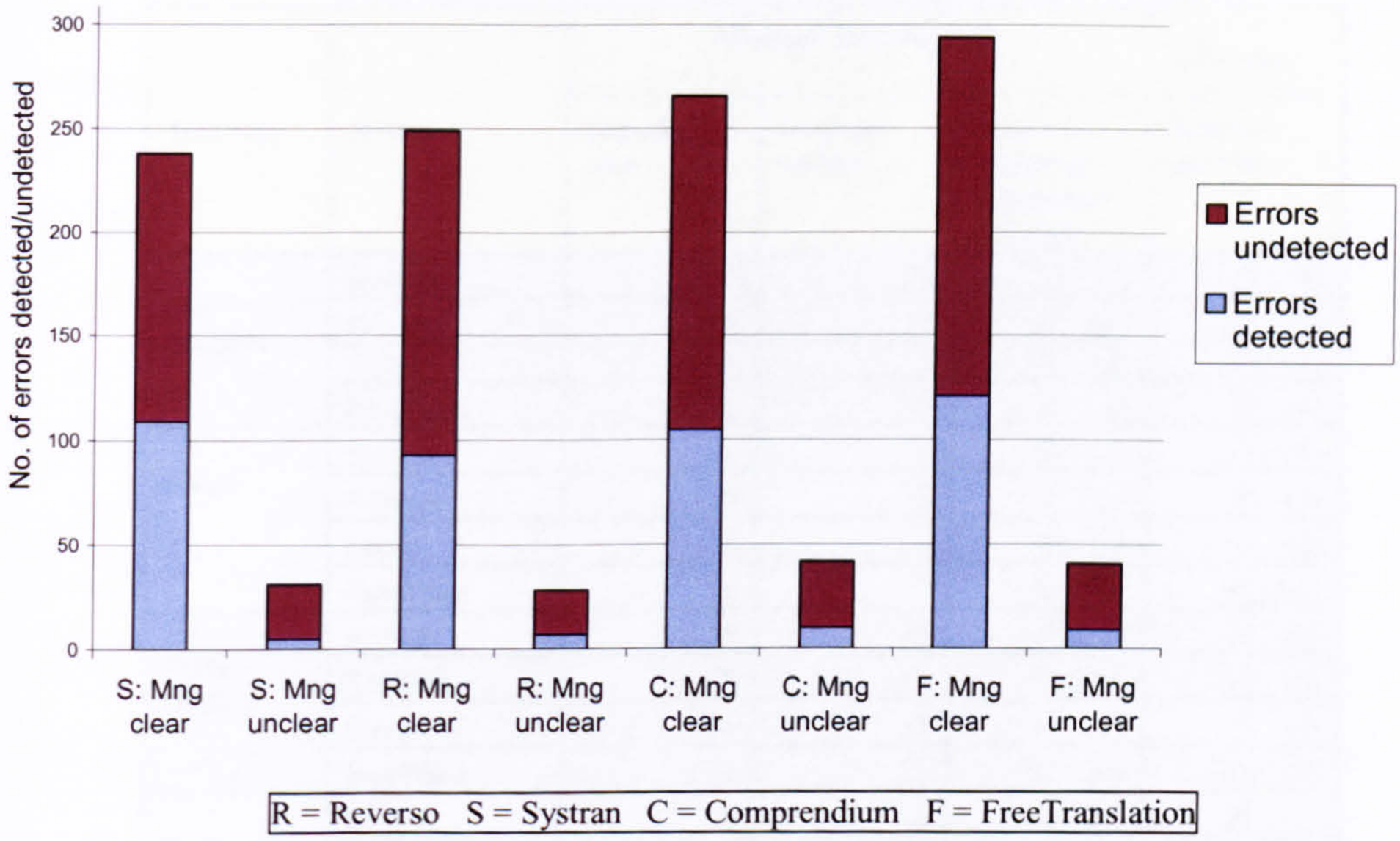


Figure 5.3: Results from automatic detection of inappropriate content words

Results showed that 428 of the 1047 errors labelled as 'meaning clear' (41%) and 31 of the 142 errors labelled as 'meaning unclear' (22%) were detected using this method, and there were no false positives. Having selected items for our rules file from a different sample of MT output (including translations by Idiomax, whose output is not included in our test sample), these findings show that there is a good deal of consistency in terms of inappropriate word combinations found in output from different MT systems. Automated error counts by system and text type are presented in Table 5.9.

Findings showed that in the case of user manuals and technical reports, the number of automatically detected errors reflects system ranks in terms of the number of manual annotations. Coinciding ranks are highlighted in Table 5.9.

Text type	System	Manual annotations			Auto score (no. of items detected) and rank
		Meaning clear	Meaning unclear	Total no. of manual annotations and rank	
User manuals	Systran	69	9	78 (3)	37 (3)
	Reverso	59	8	67 (1)	27 (1)
	Comp	64	7	71 (2)	28 (2)
	FreeTrans	71	8	79 (4)	38 (4)
FAQs	Systran	66	4	70 (1)	33 (3/4)
	Reverso	69	2	71 (2)	30 (1)
	Comp	75	4	79 (3)	31 (2)
	FreeTrans	85	5	90 (4)	33 (3/4)
Press releases	Systran	53	12	65 (1/2)	20 (2)
	Reverso	55	10	65 (1/2)	18 (1)
	Comp	64	17	81 (3)	30 (4)
	FreeTrans	70	12	82 (4)	21 (3)
Technical reports	Systran	50	6	56 (1)	24 (1)
	Reverso	66	8	74 (2)	25 (2)
	Comp	63	14	77 (3)	26 (3)
	FreeTrans	68	16	84 (4)	38 (4)
All text types	Systran	238	31	269 (1)	114 (2)
	Reverso	249	28	277 (2)	100 (1)
	Comp	266	42	308 (3)	115 (3)
	FreeTrans	294	41	335 (4)	130 (4)
TOTAL	**All systems**	**1047**	**142**	**1189**	**459**

Table 5.9: Remaining inappropriate words: manual annotations and automated scores

5.2.2.3 Correlations with human judgements

Automated scores, representing the number of inappropriate words detected using this method, were compared with human judgements for fluency and adequacy (shown in Tables 4.3 and 4.4). Just as in the case of 'outrageous' words, the number of detected 'inappropriate' words reflected the rank order of systems based on human adequacy judgements when all text types were taken into consideration. In terms of individual text types, automated scores reflected ranks according to both adequacy and fluency judgements for the technical reports and adequacy judgements for the user manuals.

Pearson's correlation co-efficient was subsequently computed to investigate correlations between automated and human scores. These are shown in Table 5.10. The strongest correlations were found when comparing the number of detected errors in the technical

reports with human judgements. Correlations for other individual text types were much weaker. Across the whole sample, automated scores correlated slightly more highly with human judgements of adequacy rather than fluency.

Texts	Value of *r* Automated scores: fluency	Value of *r* Automated scores: adequacy
User manuals	-0.3777	-0.7766
FAQs	-0.7697	-0.7078
Press releases	-0.3115	-0.0837
Tech reports	-0.9759	-0.9184
All text types	**-0.8488**	**-0.8852**

Table 5.10: Correlations: detected inappropriate words and human judgements

5.2.2.4 Conclusions

Unlike the 'outrageous' words, which were automatically detectable using the method described in 5.2.1.4, the remaining inappropriate content words were considered erroneous in their immediate context. Consequently, an updateable list of 1,490 word combinations containing 'inappropriate' words was compiled by manually selecting items from a 90,000 word sample of English MT output. Checks were made to ensure that no items were duplicated in the 'outrageous' word and 'inappropriate' word lists, to avoid any errors being detected twice. An algorithm was then used to automatically compare our smaller sample of MT output against this list of errors and to tag all matches. The algorithm could also be extended to offer error corrections to the user.

Without requiring access to the source text or to any human translation, the method was able to detect 459 (39%) of the remaining 1189 manually annotated inappropriate words in the sample and no false positives were flagged. This represents twice the number of errors automatically detected using the algorithm for outrageous words, and a much more even distribution of detected errors in each of the different text types: for inappropriate words, the highest number of errors (130) was detected in the user manuals and the lowest number (89) was found in the press releases. In contrast, the algorithm to detect outrageous words found 124 errors in the user manuals, compared with only 6 in the technical reports.

Correlations between human and automated scores were weak all cases but the technical reports (see Table 5.10). However, the number of detected errors of this kind reflected the rank order of systems in terms of human adequacy judgements for the whole sample. The detection of additional error types should be investigated to achieve stronger correlations with human scores.

5.2.3 Automatic detection of inappropriate prepositions

Prepositions are defined as inappropriate in cases where, with access only to the MT output, the wrong preposition is identified with a verb, noun or adjective, contravening grammatical rules or preferences in the given context. Based on the manual annotations in our sample, these errors appear to occur for two main reasons:

1. Inadequate prepositional rules in the MT system;
2. The acceptable use of different prepositions in French in some expressions in this domain: Eg. 'Aller **sur** la page d'accueil' is translated literally as: 'Go **on** the home page' rather than following target language conventions to generate the preferred translation: 'Go to the home page'.

Examples of errors in this category are shown in Table 5.11.

Preposition type:	**Example of MT output (Correct preposition shown in brackets)**
With verb	This function allows you to print documents **chosen in (from)** the menu. **Go on (to)** the page then click … **Delete** the text **of (from)** the page.
With noun	It is possible to forbid the **access of (to)** certain unwanted sites … … via the standard window of selection **of (in) Windows.** Sketch can open files **to (in) the Xfig format,** …
With adjective	This link is **available in (from)** the menu … It is **incomprehensible for (to)** a human. They are **incapable to (of)** using it.

Table 5.11: Examples of inappropriate prepositions

In total, 323 inappropriate prepositions were manually annotated in the 20,000 word sample. Errors were further labelled according to their association with either a verb, noun or adjective, although this information did not prove to be useful for error detection. Table 5.12 shows the number of inappropriate prepositions by system; while errors were fairly evenly distributed across systems, they still reflected the rank order (in terms of fluency) for system performance as a whole.

System	Preposition inappropriate with verb	Preposition inappropriate with noun	Preposition inappropriate with adjective	Total
Systran	20	49	4	73
Reverso	20	53	4	77
Comp	19	62	3	84
FreeTrans	23	62	4	89
All systems	**82**	**226**	**15**	**323**

Table 5.12: Inappropriate prepositions by system

5.2.3.1 Development of the detection method

As no existing methods were found for the automatic detection of inappropriate prepositions in MT output, the errors were first analysed to guide a data-driven approach. In total, 35 different words had been manually annotated in this category, as shown in Figure 5.4. A small number of these are not classed as prepositions in normal English usage, but they were found in the MT output where a preposition should have been.

apart from	besides	in connection with	opposite
around	by	in front of	since
as	during	into	through
as early as	face to face	near	to
as for	facing	next to	towards
as of	for	of	with
at	from	on	vis-a-vis
at the time of	in	on behalf of	while
beside	in comparison with	onto	

Figure 5.4: Inappropriate translations of French prepositions found through manual analysis

As in the case of inappropriate content words, many of the prepositional errors were found to occur frequently with the same or similar words. These include some of the examples in Table 5.11. Consequently, a second 'reverse' test suite (see 3.4.9) was compiled, comprising a list of inappropriate prepositions adjacent to their governing nouns, verbs or adjectives.

Word combinations were selected from the same 90,000 word sample of MT output used for the detection of inappropriate content words (see 5.2.2.1). Only the 35 items shown in Figure 5.4 were searched and analysed in the corpus to speed up the manual process. Inappropriate items were added to a rules file against which our smaller sample would be automatically compared and error matches tagged. However, many of the prepositional errors found were not added to the rules file for a number of different reasons:

1. Some words governing inappropriate prepositions were found to be very infrequent and unlikely to occur in millions of words of MT output;
2. Some prepositional errors were found in word combinations involving other types of error. For instance, some involved untranslated words ('l'accès **of** the data base') and it would not be efficient to hardwire such items;
3. Considerable distance between the inappropriate preposition and the word that governed it made some word combinations too long for inclusion (Eg. '**In** the list of computers **select** the desired computer.') The longer the entry in the rules file, the less likely its occurrence will be in a different sample of MT output. However, this problem was found to occur relatively infrequently; 84% of inappropriate prepositions were found adjacent to or within two words of their governing noun, verb or adjective.
4. Real world knowledge would be required to recognise some prepositional errors. For instance, in the phrase 'by clicking on the icon **to** the right of the toolbar' it was clear that the icon was part of the toolbar, so the preposition here should be '**on**'). However, 'to the right of' cannot be added to the rules file as this is perfectly legitimate in many circumstances.

As in the case of inappropriate content words, some of the selected word combinations were searched in the BNC to determine whether or not they are legitimate in the English

language. Some items were subsequently deleted from the rules file if they were expected to generate false positives. For items containing verbs, entries were added to ensure that all conjugations were included where necessary, and that for items involving nouns, these were entered in their singular and plural forms where appropriate. A file of rules was then formatted, with tags to be automatically inserted into the MT output after each error. The updateable file contained 572 entries. Corrections could also be added to each entry for automatic insertion into the MT output.

5.2.3.2 Results

The 20,000 word sample of MT output was automatically compared against the rules file and all found inappropriate prepositions flagged. Automatically tagged tokens and manual annotations were compared. Findings by system are shown in Figure 5.5.

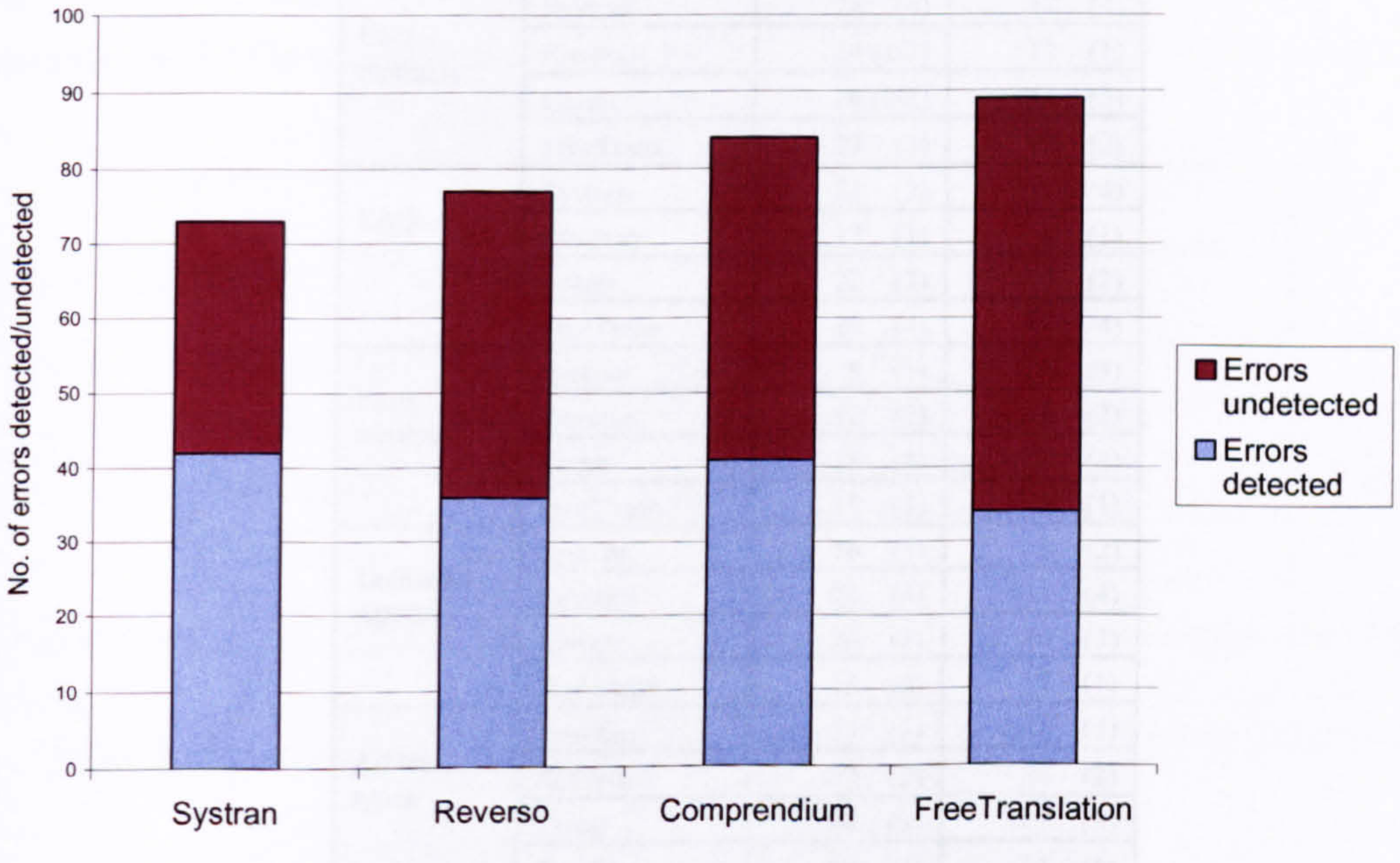


Figure 5.5: Results from automatic detection of inappropriate prepositions

Across all systems, 153 (47%) of the 323 manually annotated inappropriate prepositions were detected using this method, and there were no false positives. As with the detection of inappropriate content words, the fact that the prepositional rules file was compiled using a completely different sample of MT output (in the same domain) confirms the high level of consistency in terms of errors made by the different MT systems.

An important observation based on these initial results was that the number of automatically detected errors by system did not reflect the rank ordering of systems in terms of the number of manual annotations. In particular, in output from the weakest system, which contained the largest number of manually annotated prepositional errors, the smallest number of errors was detected automatically. The main reason for this was that output from FreeTranslation contained the highest number of all error types. This hindered the detection of inappropriate prepositions when the surrounding text was also incorrect, and in some of these cases, even the human identification of these errors was difficult. Automated error counts by system and text type are shown in Table 5.13.

Text type	System	Total no. of manual annotations and rank	Auto score (no. of items detected and rank)
User manuals	Systran	28 (4)	17 (4)
	Reverso	24 (1/2)	12 (1)
	Comp	24 (1/2)	14 (3)
	FreeTrans	27 (3)	13 (2)
FAQs	Systran	21 (2)	11 (3/4)
	Reverso	17 (1)	8 (1)
	Comp	22 (3)	9 (2)
	FreeTrans	29 (4)	11 (3/4)
Press releases	Systran	8 (1)	6 (3)
	Reverso	12 (2)	5 (2)
	Comp	18 (4)	8 (4)
	FreeTrans	15 (3)	3 (1)
Technical reports	Systran	16 (1)	8 (2)
	Reverso	24 (4)	11 (4)
	Comp	20 (3)	10 (3)
	FreeTrans	18 (2)	7 (1)
All text types	Systran	73 (1)	42 (4)
	Reverso	77 (2)	36 (2)
	Comp	84 (3)	41 (3)
	FreeTrans	89 (4)	34 (1)
TOTAL	**All systems**	**323**	**153**

Table 5.13: Inappropriate prepositions: manual annotations and automated scores

Results showed that for any text type, there was no rank correlation by system between the number of automatically detected errors and the number of manual annotations. Coinciding ranks are highlighted in Table 5.13.

5.2.3.3 Correlations with human judgements

Automated scores representing the number of inappropriate prepositions were compared with human judgements for fluency and adequacy (shown in Tables 4.3 and 4.4). As expected from findings in 5.2.3.2, when all texts were taken into consideration, no rank correlation was found. In fact, the automatic detection of inappropriate prepositions using this sample indicated that FreeTranslation was the best system. Even at the level of individual text types, there was no rank correlation between automated scores and manual annotations for either fluency or adequacy.

For the sake of completeness, Pearson's correlation co-efficient was computed between automated and human scores, even though very low values were expected. Findings are shown in Table 5.14. Bearing in mind that the perfect correlation would be -1 (the greatest number of errors representing the poorest quality), these values show the contrary in most cases. The only negative correlation found was when comparing the number of detected errors in the FAQs with human judgements.

Texts	Value of *r* Automated scores: fluency	Value of *r* Automated scores: adequacy
User manuals	0.3877	0.0704
FAQs	-0.7697	-0.7078
Press releases	0.4358	0.6062
Tech reports	0.5184	0.3934
All text types	**0.6131**	**0.6410**

Table 5.14: Correlations: detected inappropriate prepositions and human judgements

5.2.3.4 Conclusions

A list of word combinations involving inappropriate prepositions was compiled by manually selecting items from a 90,000 word sample of MT output. An algorithm was then used to automatically compare our smaller sample of output against the list of errors and to flag all matches. The program could also be extended to offer error corrections to the user. This would simply involve the addition of a corresponding correction list to the file of errors and their tags.

Without access to the source text or to any human translation, 153 (47%) of the 323 prepositional errors in the sample were detected using this method, and no false positives were flagged. This figure shows the extent to which the same errors can occur in different samples of output in the same domain. The main problem with this method, however, was that inappropriate prepositions found adjacent to other errors (such as untranslated or inappropriate words) were not detected, as these items were not incorporated in the rules file for reasons described in 5.2.3.1.

Although the number of manually annotated prepositional errors reflects the rank ordering of systems in terms of fluency, correlations between human and automated scores were extremely weak and, in many cases contradictory, whether at the level of text type or when taking the whole sample into account. In spite of this, the method is particularly useful for developers, as it automatically flags frequently occurring prepositional errors which could be a focus for system improvements. Furthermore, in conjunction with automated scores from the detection of other error types, scores from this method could still contribute to strong correlations with human judgements.

5.2.4 Automatic detection of inappropriate pronouns

'Inappropriate' pronouns are defined as mistranslations of pronouns, which are observable as incorrect without access to the source text, due to ungrammaticality or impossibility based on world knowledge. A total of 104 pronouns were manually annotated as 'inappropriate' in the 20,000 words of MT output from the four competing systems. These were initially divided into forty-four anaphoric and sixty non-anaphoric pronoun errors for subsequent analysis and error detection. The number of errors by system is shown in Table 5.15.

System	Inappropriate pronouns: anaphoric	Inappropriate pronouns: other	Total
Systran	1	13	14
Reverso	28	8	36
Comp	4	19	23
FreeTrans	11	20	31
All systems	**44**	**60**	**104**

Table 5.15: Inappropriate pronouns by system

The largest number of pronominal errors was found in output from Reverso, one of the best performing systems according to human judgements; the majority of these being incorrect anaphoric pronouns. Examples of the two kinds of pronominal errors are provided in 5.2.4.2 and 5.2.4.3.

5.2.4.1 Existing approaches to anaphoric pronoun resolution

Existing approaches to anaphoric pronoun resolution were first examined to determine whether any method might be adaptable to automatically detect incorrect anaphoric pronouns in the MT output.

An algorithm for anaphoric reference resolution typically identifies candidate referents for each pronoun, within the same and recent sentences. The most likely referent from the set of candidates is then selected, based on certain constraints such as number, person, gender and case agreement, grammatical position, syntactic constraints and preferences (eg. for a subject or object antecedent) and selectional restrictions. For example, in (1) there are five possible referents (shown in red) for the anaphoric pronouns 'he' and 'them'.

(1) **The students handed their work and textbooks to Janet to pass on to Mr. Brown. He had tried hard to motivate them this year.**

An algorithm identifying the referent of 'he' (singular, animate, male) should eliminate all candidates that are not singular ('the students', 'textbooks'), not animate ('their work') and not male ('Janet'), resulting in the selection of 'Mr. Brown'. Identifying the referent of 'them' (plural, male or female, animate or inanimate) is more complex. Candidates in the singular form can be discarded ('their work', 'Janet', 'Mr. Brown'), leaving two remaining possibilities ('the students' and 'textbooks'). Selectional restrictions must now be applied, based on real-world knowledge that can be difficult to encode. In this case, constraints placed by the verb 'motivate' on its arguments dictate that the pronoun should refer to the students rather than the textbooks, based on the verb's preference for an animate direct object, subject to exceptions.

5.2.4.1.1 Accuracy of symbolic pronoun resolution algorithms

The pronoun resolution algorithm developed by (Lappin & Leass, 1994) takes many of the above constraints into consideration, and computes a salience value for each entity. After the development of a weighting scheme using a corpus of computer training manuals, the algorithm accurately identified referents with an accuracy of 86% on unseen data in the same genre. A tree search algorithm developed by (Hobbs, 1978) searches for an appropriate antecedent noun phrase among syntactic representations of sentences. This method achieved an accuracy of 88.3% on three different texts, increasing to 91.7% with the addition of selectional restrictions. Centering algorithms such as BFP (Brennan, Friedman and Pollard, 1987) compute preferred referents according to their relations with forward and backward-looking centres in adjacent sentences, applying rules and salience factors such as grammatical role, recency and repeated mentions of candidate referents. Walker's evaluation of the BFP algorithm (Walker, 1989) revealed an accuracy of 77.6% on texts from three different genres. Her evaluation of Hobbs' algorithm with the same texts returned an accuracy of 81.8%. Strube's S-list (salience list) algorithm, motivated by BFP, correctly resolved 85% of pronouns in short stories and news articles, while his evaluation of the BFP algorithm resolved only 76% of pronouns in the same texts. More recently, Tetrault's Left-Right Centering algorithm achieved most success with fictional texts, correctly resolving 81.1% of pronouns.

5.2.4.1.2 Accuracy of machine learning methods

Machine learning approaches to coreference resolution first involve the automatic identification of noun phrases (NPs) in a document. A learning algorithm is used to train a classifier which then compares each NP to each preceding NP and returns a number between 0 and 1, indicating the likelihood of the two NPs being coreferent. In 2004, Ng conducted an evaluation of the learning-based coreference system developed by (Ng and Cardie, 2002) and enhanced with anaphoricity information (Ng, 2004). With 422 news texts from three sources, the maximum F-measure reported for coreference resolution was 69.1 in Broadcast News texts.

An experiment by (Strube and Müller, 2003) involved the addition of specifically designed features to a decision tree-based learning approach to resolve pronouns in a corpus of spoken switchboard dialogues. Pronoun resolution in spoken dialogue presents particular

difficulties, mainly because many pronouns have non-NP antecedents or no antecedent at all: (Eckert and Strube, 2000) found that 22% of pronouns in a set of switchboard dialogues had non-NP antecedents, and 33% had no antecedent. Further hindrances to pronoun resolution can be due to disfluencies, abandoned utterances and interruptions. (Strube and Müller, 2003) report that for many pronouns in spoken dialogue, the antecedent cannot even be determined by humans. Their evaluation of this approach to pronoun resolution in spoken dialogue reported a precision of 61.71% for all pronouns and 80% for third person masculine and feminine pronouns only. Recall, however, was very low in each case, at 36.48% and 9.6% respectively.

5.2.4.1.3 Suitability of existing algorithms for error detection

Attempting to modify any of the above approaches to the automatic detection of incorrect anaphoric pronoun usage in MT output would be inappropriate for the following reasons:

1. Such algorithms are designed to resolve pronouns in correct usage. Modifying an existing method in an attempt to detect *incorrect* anaphoric pronouns is not feasible, as pronoun error detection would only be successful in cases where no compatible referent was found. This would be highly unlikely unless search constraints were applied; as such an algorithm would continue to search until a candidate referent was found. Adding to this problem is the fact that number and gender constraints can also be violated, as in example (2), leading to incorrect pronoun resolution even when pronouns are used legally.

(2) He took his car to the garage. *They* asked him to pick it up on Tuesday.
(Katja Markert, personal communication, 2005)

2. The algorithms work on the assumption that written input is well-formed: the machine-translated texts under analysis here contain an average of 4.8 incorrect items per sentence. Error types that would particularly impede anaphoric resolution, include:
    a. inappropriate, untranslated, and unnecessary nouns, named entities and noun strings (of up to seven words), all of which may be candidate referents
    b. omitted words

c. incorrect prepositions, which can confuse relationships between verbs and their objects.
d. inappropriate verbs, which may place quite different constraints on their arguments when compared with their correct translations
e. untranslated and (in rare instances) omitted verbs

Some of the problems presented by ill-formed input are comparable to those encountered when working with spoken dialogue. However, (Strube and Müller, 2003) modified their data by excluding disfluencies and abandoned utterances from analysis to improve the accuracy of their approach. Such modifications are not appropriate for this research, as disfluencies and their detection are the focus.

3. The frequency of incorrect anaphoric pronouns in the corpus represents only around 2% of all errors under consideration for automated detection, making such an investigation far beyond the scope of this thesis.

5.2.4.2 Development of a detection method for inappropriate anaphoric pronouns

The forty-four anaphoric pronominal errors were analysed in order to find an alternative detection method. These errors were found to involve eight pronoun forms: **he, she, him, his, her, himself, it, its.** With access only to the machine translation output, it was clear that all of these errors were due to multiple translation possibilities for third person pronouns from French into English, as shown in Table 5.16.

French pronoun	Possible English translation
il	he, it
elle	she, it
le	him, it, the
la	her, it, the
son	his, her, its
sa	his, her, its
ses	his, her, its
lui-meme	himself, itself
se (with reflexive verb)	himself, herself, itself

Table 5.16: English translations of French pronouns

The attribution of gender to inanimate objects in French caused 'il' or 'elle', for example, to incorrectly translate into 'he' or 'she' rather than the required 'it' in some instances. The example in (3) from Reverso ProMT shows two such annotated errors, caused by the fact that 'computer' is a masculine noun in French.

(3) **6.2.1 How to define slots of use**
Slots can be defined a day for global duration or a day on an authorized slot. As soon as the computer is outside the track or exhausted his duration of authorized connection, he has no more access to internet.

It should be noted here that although these observations are specific to the French-English language pair, similar errors can occur in output from other gender-based source languages, such as Italian, German and Spanish. The automatic detection of pronominal errors in translations from these languages is described in 6.4.

The sample of MT output was subsequently analysed to establish the raw frequency and error frequency of each of the eight pronouns under investigation. Only two anaphoric errors were annotated in 225 instances of the pronoun 'it', and one error was found in 32 occurrences of 'its'. However, of the six remaining pronouns, 41 of the 49 occurrences had been annotated as errors, as shown in Figure 5.6.

Findings showed that if these six words were automatically tagged as errors, with no additional rules or complex linguistic knowledge, 41 of the 44 annotated anaphoric pronoun errors would be detected, with eight false positives, achieving 83.6% precision and 93.1% recall.

Table 5.17 shows the frequency of the six pronouns and the number of annotated errors found in the 5,000 words of MT output for each text type.

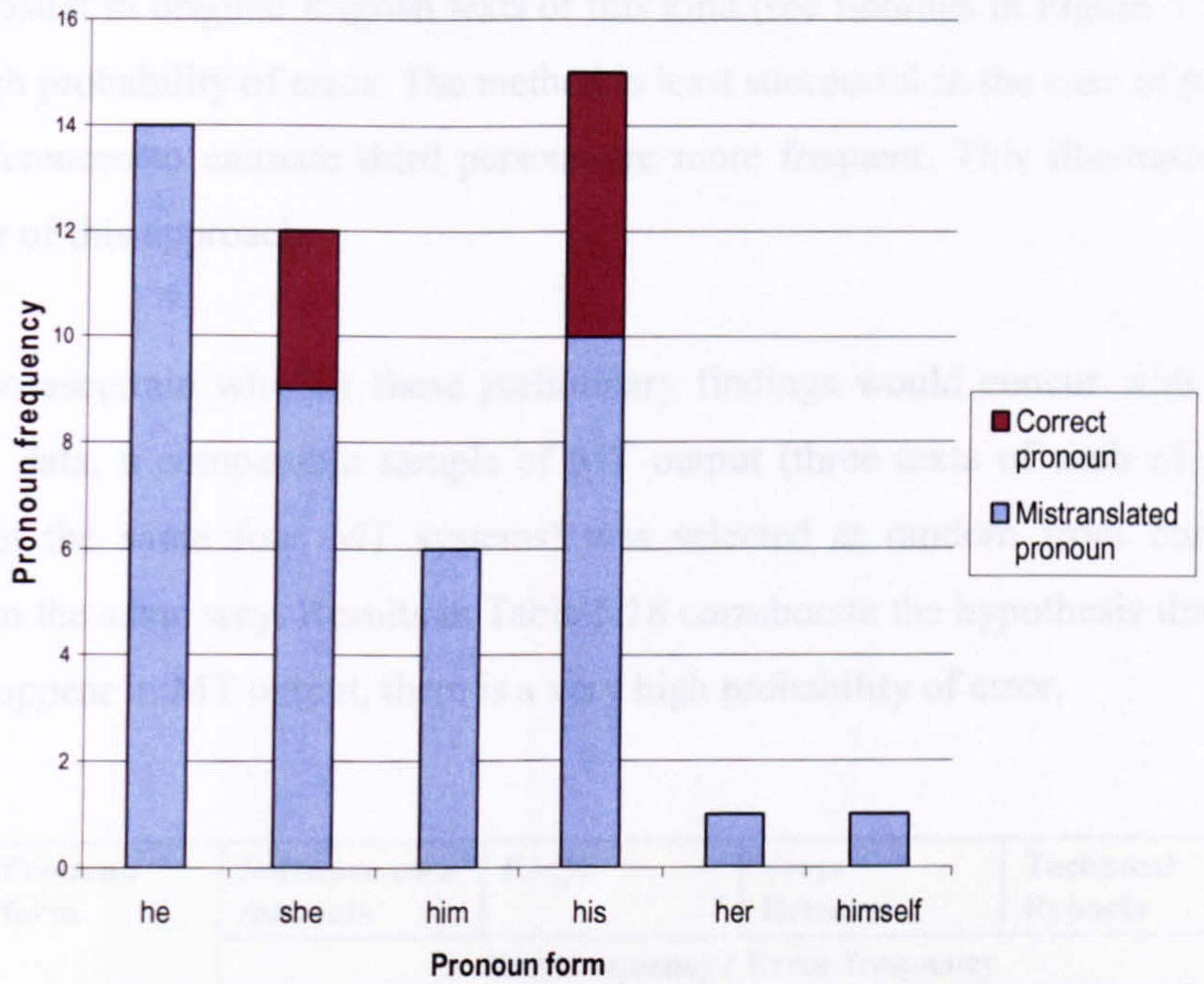


Figure 5.6: Pronoun frequency and annotated pronominal errors

Pronoun form	Software user manuals	FAQs	Press Releases	Technical Reports
	Raw frequency / Error frequency			
he	6 / 6	3 / 3	2 / 2	3 / 3
she	3 / 3	1 / 1	2 / 2	6 / 3
him	3 / 3	0 / 0	0 / 0	3 / 3
his	4 / 4	2 / 0	4 / 1	5 / 5
her	0 / 0	0 / 0	0 / 0	1 / 1
himself	0 / 0	0 / 0	0 / 0	1 / 1
Total	**16 / 16**	**6 / 4**	**8 / 5**	**19 / 16**
Precision	**100%**	**66.6%**	**62.5%**	**84.2%**
Recall	**100%**	**100%**	**100%**	**100%**

Table 5.17: Frequency of pronouns and annotated errors by text type

Results from this sample show that automatically flagging these six pronouns as errors is most reliable in the case of user manuals, where all occurrences are actual errors. This is due to the fact that such technical texts, which instruct the reader directly, very rarely make reference to animate third persons. This makes the use of the pronouns shown in Table 5.17

highly unusual in original English texts of this kind (see findings in Figure 5.7), indicating a very high probability of error. The method is least successful in the case of press releases, where references to animate third persons are more frequent. This illustrates the limited portability of this approach.

In order to ascertain whether these preliminary findings would concur with results from additional data, a comparable sample of MT output (three texts of each of the four text types from the same four MT systems) was selected at random from our corpus and analysed in the same way. Results in Table 5.18 corroborate the hypothesis that when these pronouns appear in MT output, there is a very high probability of error.

Pronoun form	Software user manuals	FAQs	Press Releases	Technical Reports
	Raw frequency / Error frequency			
he	1 / 1	1 / 1	1 / 1	0 / 0
she	2 / 2	0 / 0	0 / 0	0 / 0
him	0 / 0	2 / 2	1 / 1	0 / 0
his	3 / 3	0 / 0	11 / 11	5 / 5
her	1 / 1	0 / 0	1 / 1	0 / 0
himself	1 / 1	0 / 0	0 / 0	0 / 0
Total	**8 / 8**	**3 / 3**	**14 / 14**	**5 / 5**
Precision	**100%**	**100%**	**100%**	**100%**
Recall	**100%**	**100%**	**100%**	**100%**

Table 5.18: Frequency of pronouns and annotated errors in comparable sample

5.2.4.2.1 Analysis of correct pronoun usage in original English texts

The frequency of these six pronouns in original English texts was subsequently computed, to establish the number of occurrences expected in correct usage. Representing each of the four text types, a one million word ad hoc web corpus was compiled and analysed. Figure 5.7 shows the number of words in which we would expect to find a single occurrence of each pronoun. NB. Figures above 120,000 words are not shown.

The most frequent pronoun found in a single text type was 'his' in the press releases, with 260 occurrences (1 in 3,846 words). Across all text types, 'himself' was the least frequent,

with 8 occurrences in the FAQs (1 in 125,000 words), and no instances in the other three text types.

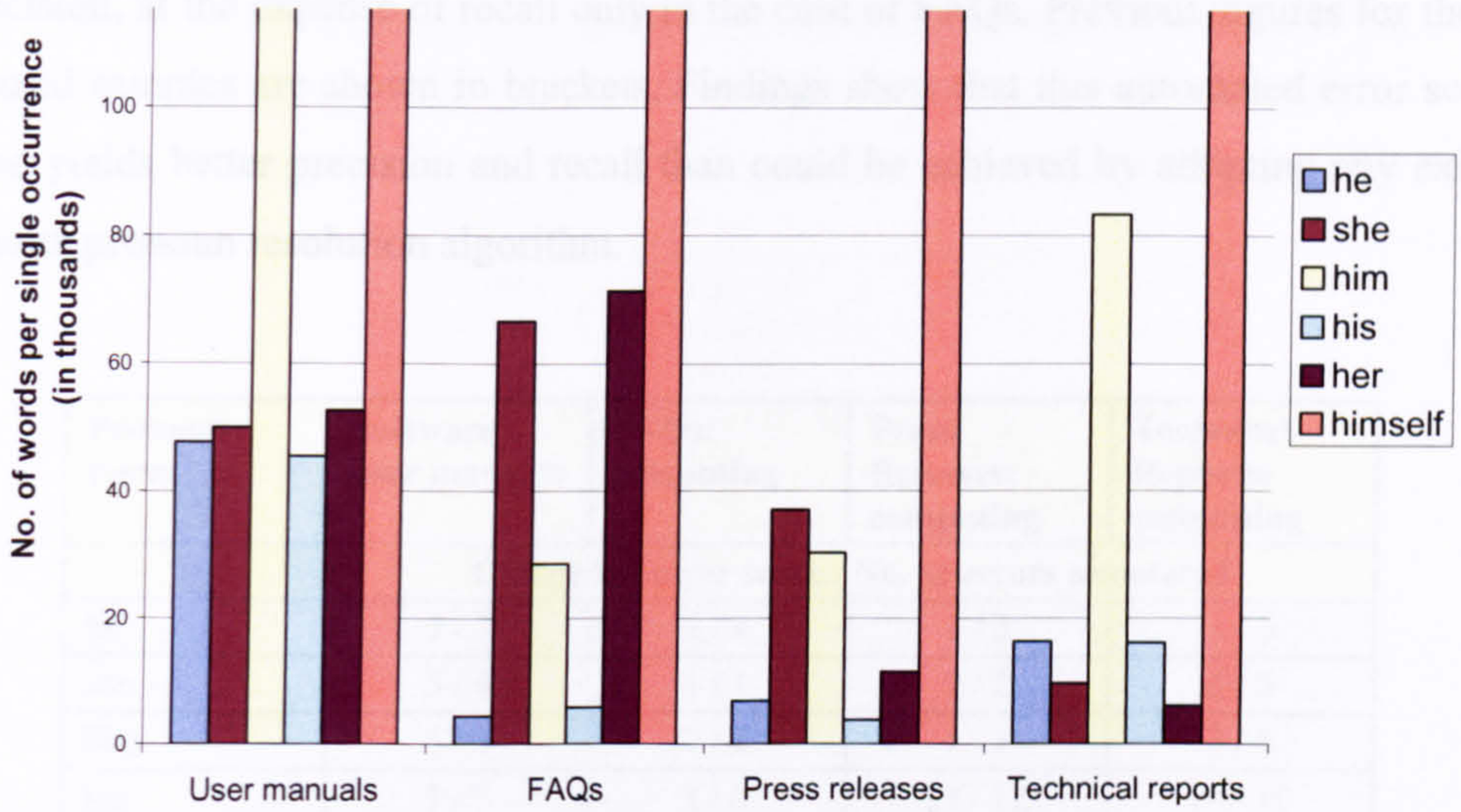


Figure 5.7: Number of words per single occurrence of pronouns in original English texts

5.2.4.2.2 Adjusting automated error scores

The frequencies illustrated in Figure 5.7 can be used to adjust the automated MT error score (based on the raw frequency of each pronoun) by subtracting the estimated frequency of correct usage. The following mathematical formula was devised to represent this calculation:

$$e_i = f(Pr_i) - \frac{n}{R_i}$$

where, for any sample of MT output, the error score *e* for any one of the six pronouns is equal to the frequency of that pronoun *Pr* in the sample, minus the number of words *n* in the sample divided by the false positive ratio *R*; this being equal to the number of words in which we would expect to find a single correct occurrence of that pronoun in the given text type.

The pronoun and error frequencies shown in Tables 5.17 and 5.18 were combined to provide a larger sample of MT output on which to test this formula. Based on 10,000 words per text type, six of the computed error scores required adjustment. The amended figures are highlighted in Table 5.19. Applying this equation leads to a considerable improvement in precision, at the expense of recall only in the case of FAQs. Previous figures for the two combined samples are shown in brackets. Findings show that this automated error scoring method yields better precision and recall than could be achieved by adopting any existing anaphoric pronoun resolution algorithm.

Pronoun Form	**Software user manuals**	**FAQs: computing**	**Press Releases: computing**	**Technical Reports: computing**
	Computed error score / No. of errors annotated			
he	7 / 7	**2** / 4	**2** / 3	3 / 3
she	5 / 5	1 / 1	2 / 2	**5** / 3
him	3 / 3	2 / 2	1 / 1	3 / 3
his	7 / 7	**1** / 0	**13** / 12	10 / 10
her	1 / 1	0 / 0	1 / 1	**0** / 1
himself	1 / 1	0 / 0	0 / 0	1 / 1
Total	**24 / 24**	**6 / 7**	**19/19**	**22/21**
Precision	**100% (100%)**	**100% (77.8%)**	**100% (86.4%)**	**95.4% (87.5%)**
Recall	**100% (100%)**	**85.7% (100%)**	**100% (100%)**	**100% (100%)**

Table 5.19: Pronominal error scores and manual annotations by text type

A simple program was written to automatically flag these pronouns as errors. As with the previous error types investigated, error corrections could easily be added to the rules file if required, to be automatically inserted into the MT output at the appropriate points of error.

5.2.4.3 Development of a detection method for inappropriate subject and relative pronouns

The remaining sixty pronominal errors were sub-divided into four categories according to how error detection might take place:

- incorrect subject pronouns (16 errors)

- incorrect relative pronouns (15)
- stylistic errors (13)
- pronouns in erroneous context (16)

Table 5.20 shows the frequencies of incorrect subject and relative pronouns in the sample, with suggested correct translations.

MT error	Preferred translation if known	Total
Subject pronoun errors		
one	you	**14**
us	we	**1**
them	they	**1**
Relative pronoun and related errors		
all what	all that	**1**
this that	which	**3**
, what	which	**7**
that it that	that what	**1**
is indeed for it	is for this reason	**1**
is well for that	is for that reason	**2**
Total		**31**

Table 5.20: Incorrect subject and relative pronouns

5.2.4.3.1 Subject pronoun errors

Three pronoun forms were manually annotated as incorrect. The use of 'us' and 'them' as subject pronouns is clearly ungrammatical. The use of 'one', however, is a case of inappropriate style within particular genres: no instances of 'one' as a subject pronoun were found in our three-million word corpus of English user manuals and FAQs. Examples of this error in the MT output are shown in (5).

(5) **One** can therefore by simple mice click …
… with the aid of a CD-ROM that **one** will put …

These errors are due to the frequent use of 'on' in French, which may translate into 'you', 'people', 'they', 'we', or (in rare instances) 'one'. In the case of user manuals and FAQs,

the required translation in all cases is 'you'. However, for press releases and technical reports the use of 'one' as a subject pronoun is more widely acceptable and is not considered as an error in these text types.

The pronoun error detection algorithm was extended to flag and offer a correction for 'Us' and 'Them' when found at the start of a sentence (as in our MT output), and for 'one' when followed by the small number of verbs it was found to govern in the user manuals and FAQs ('can', 'cannot', 'must', 'will'). NB. Although a part-of-speech tagger could be used to investigate the potential of a more robust detection method here, this was deemed impractical due to the trivial nature and number of this kind or error.

5.2.4.3.2 Relative pronouns and related errors

Even without access to the source text, it was clear that errors involving relative pronouns were mainly due to the literal translation of individual words, in cases where particular combinations of words in the source language would require a different translation. One example is shown in (6).

(6) ce = this/it
que = that/which
ce que = which/what
ce que = this that (MT)

The relative pronoun errors in Table 5.20 were subsequently searched in the written part of the BNC to determine whether circumstances exist in which these word (and punctuation) combinations can be correctly used in English. In most cases, no occurrences were found, indicating that these items could be automatically flagged as errors. The rare exceptions were instances of ', what', appearing either as an error or in works of fiction, and 'all what', which is acceptable in particular, rare circumstances. Simple exception rules can ensure that these correct instances are not identified as errors.

5.2.4.3.3 Stylistic pronominal errors

The thirteen manually annotated stylistic errors involved eight word combinations. The annotator's intuitions were tested by searching these items in the BNC, along with the

annotator's preferred translation, to determine the frequency in each case. Findings are shown in Table 5.21. Examples of stylistic pronoun errors in context are shown in (7).

MT output	Frequency in written BNC	Annotator's preference	Frequency in written BNC
it serves	144	this serves	30
it is used	335	this is used	44
which you	2,254	that you	11,541
Does that	621	Does it	1,960
That will	4,240	It will	15,280
that has	4,497	it has	18,378
that explains	64	which explains	145
resulted from it	2	resulted from this	14

Table 5.21: Errors annotated due to stylistic preference, with frequencies in the BNC

(7) Sketch stores all the operations **which[that]** you carry out

… bad communication resulted from **it[this]** between the different people

The annotator's intuitions were supported by the fact that in six of the eight cases, her preferred translations were more frequent than the machine translations in the written part of the BNC. However, these annotated errors are not grammatically incorrect, nor do they have any effect on the reader's understanding of the text. Although the annotator's preferences appear, in most cases, to be more common, the word combinations found in the MT output are still comparatively frequent in well-formed English. For these reasons, a decision was taken not to include these errors in the detection algorithm.

5.2.4.3.4 Inappropriate pronouns in erroneous context

The automatic detection of the remaining sixteen errors was not considered, due to the additional errors in the surrounding text. Two examples are shown in (8).

(8) … the last documents on which **these** worked you.

… if stocked the news limited **themselves** in the name of account …

As noted in the case of inappropriate content words and prepositions, rare and erroneous word combinations such as these should not be added to the rules file for reasons stated in 5.2.2.1 and 5.2.3.1.

5.2.4.4 Results

An algorithm was used to automatically detect and annotate the anaphoric, subject and relative pronouns described above in our sample of MT output. Results by system are shown in Figure 5.8.

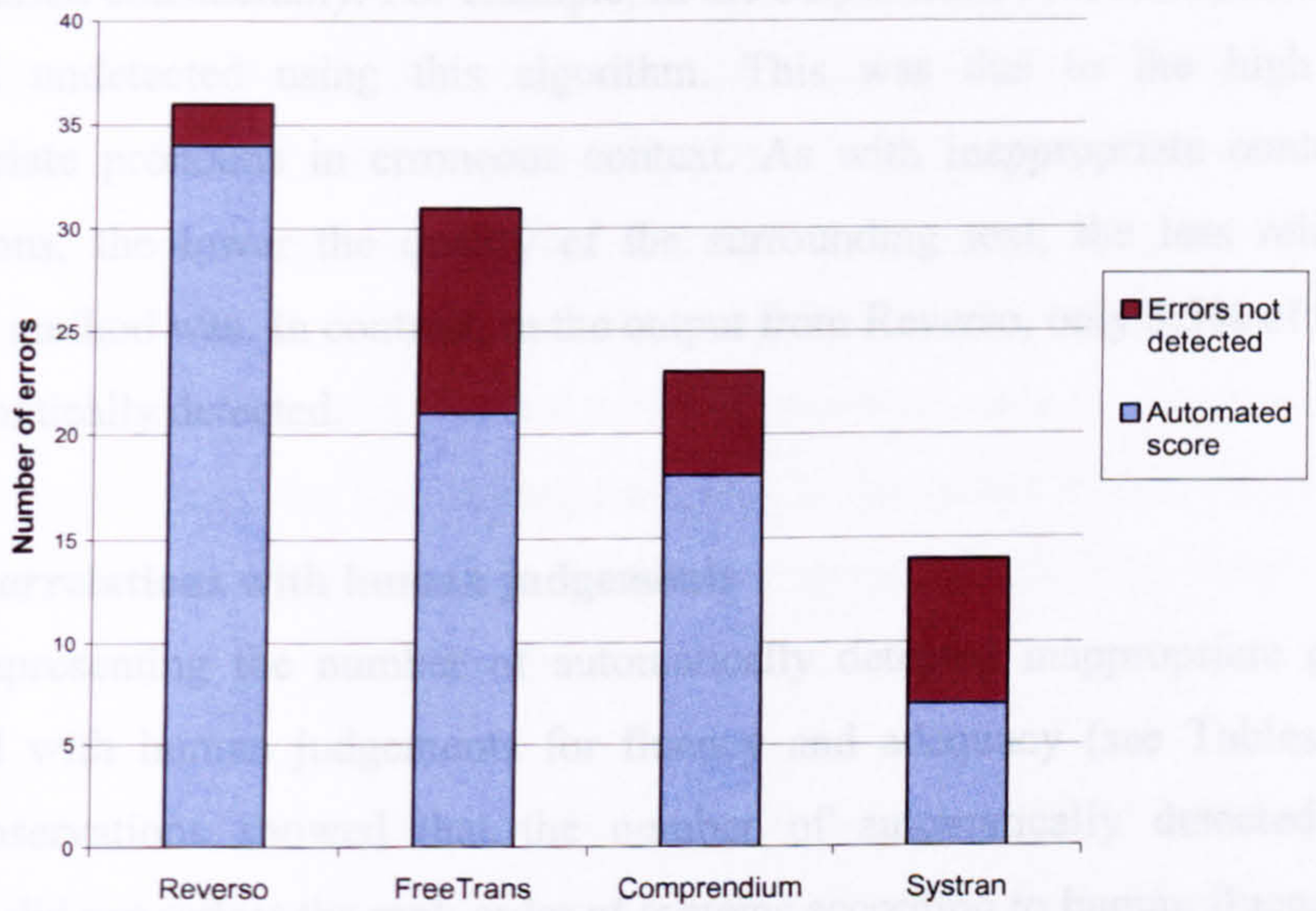


Figure 5.8: Inappropriate pronouns: detection results by system

Of the 44 manually annotated anaphoric pronoun errors, 3 were not detected (it/its) and there were 8 false positives, giving a total automated error count of 49. Of the remaining errors, 31 out of 60 were automatically detected, with no false positives. Automated scores and manual annotations are shown for all text types and systems in Table 5.22.

Text type	Systran		Reverso		Comprendium		FreeTranslation		Total	
	Auto score	Actual errors	Auto score	Actual errors	Auto score	Actual errors	Auto score	Actual errors	Auto score	Actual errors
Manuals	1	3	13	13	1	1	5	7	**20**	**24**
FAQs	5	4	2	2	9	11	5	8	**21**	**25**
Press	1	0	4	6	2	2	4	6	**11**	**14**
Reports	0	7	15	15	6	9	7	10	**28**	**41**
Total	**7**	**14**	**34**	**36**	**18**	**23**	**21**	**31**	**80**	**104**

Table 5.22: Distribution of errors across text types and MT systems

Scores based on the number of automatically detected inappropriate pronouns were found to predict the ranking of the four MT systems in terms of the number of manually annotated errors, as can be seen in Figure 5.8. This still held true at the level of individual text types, although for user manuals and FAQs two systems with adjacent ranks in terms of actual errors produced equal automated scores.

An important observation was that the percentage of errors detected in the output from each system varied considerably. For example, in the output from FreeTranslation 32% of errors remained undetected using this algorithm. This was due to the high frequency of inappropriate pronouns in erroneous context. As with inappropriate content words and prepositions, the lower the quality of the surrounding text, the less reliable the error detection method was. In contrast, in the output from Reverso, only 5.5% of the errors were not automatically detected.

5.2.4.5 Correlations with human judgements

Scores representing the number of automatically detected inappropriate pronouns were compared with human judgements for fluency and adequacy (see Tables 4.3 and 4.4). Initial observations showed that the number of automatically detected inappropriate pronouns did not reflect the rank order of systems according to human fluency or adequacy judgements. No rank correlations were found, whether taking all system output or individual text types into consideration. Pearson's correlation co-efficient was also computed to compare automated scores with human judgements; findings are shown in Table 5.23.

Texts	Value of *r* Automated scores: fluency	Value of *r* Automated scores: adequacy
User manuals	-0.1606	0.2247
FAQs	-0.6097	-0.3325
Press releases	-0.5969	-0.7758
Tech reports	-0.1677	-0.2182
All text types	**-0.0797**	**-0.0208**

Table 5.23: Correlations: inappropriate pronouns and human judgements

Values were very low, due to two main reasons:

1) The Reverso output contains the largest number of inappropriate pronouns, despite this being one of the top-performing systems in the human evaluations;
2) The output from FreeTranslation contains a large number of incorrect pronouns, which are undetectable due to the erroneous text surrounding them.

5.2.4.6 Conclusions

An algorithm was developed to detect inappropriate pronouns in French-English MT output using a list of rules to identify particular words and word combinations. All rules were developed after detailed analyses of the BNC and the use of four ad hoc web corpora, each of one million words, representing the four text types in our sample of MT output. The method identified 72 of the 104 manually annotated incorrect pronouns, with eight false positives, without access to the source text and, therefore, akin to a native English speaker with no knowledge of the source language. Due to the nature of the algorithm, additional rules could be added at any time to detect other errors of this kind found in English MT output, regardless of the source language.

There was no correlation between the number of automatically detected inappropriate pronouns and human quality judgements of the texts as a whole. However, the automatic identification of incorrect pronouns can still be useful for developers and post-editors. Furthermore, in combination with scores from the detection of other error types, the frequency of incorrect pronouns could contribute to the refinement of automated scores for improved correlations with human judgements.

5.2.5 Automatic detection of untranslated words and acronyms

Untranslated words and incorrect acronyms are considered together here, as the same approach was used for their automatic detection. For the purpose of this work, in which error detection does not rely on the source text, untranslated items are defined as words or acronyms which are not legitimate in the English language and are, therefore, detectable by looking at the MT output in isolation.

Untranslated words occur when tokens in the source text are not found in the system dictionaries. These tokens remain in the source language and are inserted into the target text. Untranslated words can also occur as a result of incorrect tokenisation; we assume that this is the reason why a large number of nouns beginning with vowels, along with their determiners, were left untranslated by FreeTranslation (eg. l'accès, l'ordinateur, l'auteure).

A total of 168 words were manually annotated as 'untranslated' in the corpus of MT output, without access to the source text. In addition, 35 incorrect acronyms were manually tagged. In most cases these were known to be untranslated, although there were instances of mistranslations when the acronym in French was also a legitimate word. Examples of this are provided in Table 5.27. The number of untranslated words and acronyms found in output from each MT system is shown in Table 5.24, FreeTranslation having produced 53% of errors of this type.

Text type	Systran			Reverso			Comprendium			FreeTrans			TOTAL
	Untrans	Acronyms	Total	Untrans	Acronyms	Total	Untrans	Acronyms	Total	Untrans	Acronyms	Total	
Manuals	9	3	12	2	5	7	5	5	10	36	5	41	70
FAQs	10	1	11	6	0	6	9	2	11	15	0	15	43
Press	4	0	4	2	0	2	12	0	12	28	0	28	46
Reports	4	3	7	1	6	7	4	2	6	21	3	24	44
Total	27	7	34	11	11	22	30	9	39	100	8	108	203

Table 5.24: Untranslated words and acronyms by system

5.2.5.1 Existing methods for the identification of untranslated words and acronyms

To date, no published work has been found on the automatic detection of untranslated words or acronyms in MT output. However, many commercial MT engines, including the three used for this research, offer the option to highlight all words in the target text that were not found in the system dictionaries, and provide a frequency list of these tokens. The main purpose of this function is to enable the user to quickly identify terms to add to a User Dictionary for a subsequent improved translation. However, there are three main reasons why this feature cannot be exploited for this research:

1. A word marked in the target text as 'Not Found' in the system dictionaries is often not considered by the reader to be untranslated. For instance, in the 5,000 words of French-English output from Systran, 34 tokens were manually annotated as untranslated words or acronyms, without access to the source text. However, 137 tokens were flagged as 'Not Found' in the Systran dictionaries. Of these, 66 were named entities and 19 were acronyms or abbreviations (eg. file extensions and product codes), all of which needed to remain the same in the target text. A further 32 tokens were English computing-related terms, which had been used in the French source texts (eg. setup, browsers, online). The use of English terminology is common in the technology domain in French, yet these words did not appear as French terms with English 'translations' in the system dictionaries.

   Similar findings were presented by (Reeder and Loehr, 1998). Following an analysis of output from two commercial MT systems - Systran and Globalink – they found that 45% of the marked tokens were email addresses, words already in the target language and acronyms.

   In order to reduce the number of words flagged as 'Not Found', the user has the option to mark selected source text items as 'Do not translate'. However, this can be laborious and time-consuming, and contradicts the purpose of finding a fully automatic method for detecting untranslated words.

2. The annotation of words not found in system dictionaries is unreliable. For example, of the 34 tokens manually annotated as 'untranslated' in the Systran output, only 20 were flagged as 'Not Found' in the system dictionaries. Based on a black box analysis of the MT output alone, reasons for this remain unexplained.

3. The annotation of not found words is not available for online MT and some commercial systems, limiting the portability of any method that relies on this function.

5.2.5.2 Development of the detection method

The aim was to find a method for the automatic detection of (a) words considered by the reader of a machine translation to be untranslated and (b) acronyms considered to be incorrect. This could be achieved by detecting French words in the English output by, for example, automatically comparing the machine translations against a French lexicon and flagging all matches found. However, this would be inappropriate for two main reasons:

1. The intention was to devise an error detection method that could be ultimately extendible to any language pair in which the target language is English. The above approach would only be usable when the source language is French. For other source languages, additional word lists would be required, making the method too 'resource-heavy'.

2. Due to the large number of words that occur in both French and English (with or without the same meaning) automatically identifying all legitimate French words in a French-English machine translation would flag a large number of false positives.

It was considered more appropriate to automatically compare the MT output against an English lexicon to flag all tokens not found:

1. This method could be used to detect untranslated words in English MT output from any source language;

2. Although some untranslated words would not be detected if they are legitimate in both English and French (or any other source language), these could be flagged with some other error type if their use is inappropriate in a given context. For example, the untranslated French word 'moult' was detected as an 'outrageous' word in 5.2. For this research we are only concerned with the target text and assume no knowledge of the source language. With this in mind, such an automated method would identify untranslated words in the same way as a native English speaker with no knowledge of French.

An unsupervised algorithm was developed to compare the MT output against the English lexicon, described in 5.2.1.1, and to flag all tokens not found in the lexicon as 'untranslated'. This was first tested with the output from all four systems, with no additional rules. All tagged tokens in the returned text were then analysed, compared with the manual annotations, and used as a basis for the addition of rules to yield the most accurate results.

Initial tests showed that a small number of the manually annotated untranslated words from the sample were not detected using the above algorithm and lexicon. These French tokens were subsequently found in the English lexicon and deleted. Future work will involve taking measures to delete all foreign words from the lexicon, excluding those commonly-used in English. The algorithm and reasoning behind the various stages are summarised below.

(1) Convert a copy of the MT output to lower case, in accordance with the English lexicon;

(2) Remove all tokens that do not require analysis (ie. non-candidate untranslated words) to ensure correct tokenisation. These 'non-words' were found to contain various characters, as shown in Table 5.25, where potential problems are explained.

(3) Replace all hyphens in the MT output with a space: the decision was taken to treat components of hyphenated tokens separately, as in preliminary tests a number of these were incorrectly flagged as 'untranslated'. It was clear that some MT engines had separated hyphenated words not found in the system dictionaries, found translations for their component parts, then reconstructed them by replacing the hyphen. For example, translations proposed for 'copie-collant' (copy-pasting) were 'copy-stocking', 'copy-tights' and 'copy-sticking'. In the first two instances, the translations resulted from the incorrect parsing of 'collant' as a noun (tights or stocking) instead of a verb (coller – to stick or paste). Initially, our algorithm tagged all of these as 'untranslated', as they were not found in the English lexicon. However, by replacing hyphens with spaces, the constituent parts are identified in the lexicon, and are therefore not flagged. Of course, the above words should be automatically tagged as different kinds of errors.

(4) Tokenise the MT output using the Natural Language Toolkit (NLTK) Tokeniser (Loper and Bird, 2005)[4]. Tokens are split at apostrophes so that when items such as *l'information* or *d'accès* are analysed, the determiner/preposition (here: *l'* or *d'*) and the noun are analysed individually. Here, for example, we wish to identify three untranslated words: *l'*, *d'* and *accès*.

[4] Freely downloadable from http://nltk.sourceforge.net/

Characters	Example of token	Reason for exclusion	Rule
( ) { } [ ] * ! ?	Colo(u)r	Subsequent incorrect tokenisation: colo – u – r (3 tokens)	Disregard any token with an opening bracket anywhere BUT at the beginning of the token. Disregard any token with * ! ? or a closing bracket, anywhere BUT at the end of the token
_ + = | & ~ # ^ % £ $ °	L&H	Tokens containing these symbols will not appear in the lexicon. Some will cause tokenisation problems (as above)	Disregard all tokens containing any of these symbols
http www	Websites	Tokenisation will divide URLs at full stops, creating nonsense 'words' liable to be flagged as 'untranslated'	Disregard all tokens containing these character strings
@	Email addresses	Tokenisation will divide email addresses at full stops, creating nonsense 'words'. Characters surrounding the @ sign will be considered as words - eg. debe@comp	Disregard all tokens containing the @ symbol
0-9	Times, dates, ordinal numbers, product codes, section numbers, ranges, percentages, etc.	It would be impossible to add all combinations of these to the lexicon. Disregarding numbers from analysis will make the method more computationally tractable	Disregard all tokens containing numbers
. (full stop)	newfile.txt	Tokenisation will divide file names at full stops, often creating nonsense words. File names should remain untranslated	Disregard all tokens containing full stops, unless at the beginning (file extensions) or end of a token.
More than 2 consecutive occurrences of the same letter	xxx	These character strings are sometimes used to denote command line arguments. They were also found to be used as dividers between sections of text	Disregard all character strings containing more than two consecutive occurrences of the same letter

Table 5.25: Characters found in 'non-words' in the MT output

(5) Create a unique word list of all remaining tokens from the file

(6) Compare the unique word list against the English lexicon

(7) For any token not found in the lexicon, flag all occurrences of that token (regardless of case) in the *original* MT output file, which is not case-normalised.

(8) Additional rule 1: detect untranslated ordinal numbers. As all tokens containing numbers were earlier discarded from analysis, untranslated ordinal numbers (eg. 1er/1ère, 2e/2ième) would remain undetected. Therefore, an additional rule now detects and flags these, by matching tokens in the MT output against a

list of French ordinal numbers. Of course, the list needs to be extended when working with MT from other source languages.

(9) Additional rule 2: detect untranslated single letter words. The algorithm described above does not allow for the detection of many untranslated single letter words. This is because all 26 letters of the English alphabet appear as individual entries in the lexicon; if they did not, each single letter (eg. section lettering) would be flagged as an untranslated word. There are twelve single letter words (including shortened forms with apostrophes) in common usage in French, of which ten can be identified as 'untranslated' by adding further rules to the algorithm. These tokens and rules for their detection are shown in Table 5.26.

Token	Detection rule
à ô	Detected by initial algorithm, as tokens not found in English lexicon
c' d' j' l' m' n' s' t'	Tag as 'untranslated' when immediately followed by any vowel, *h* or *y* (the only legal combinations in French). This prevents the incorrect flagging of tokens before a single closing quotation mark ('us' and 'them'), possessives (a friend's computer) and plurals (LED's)
y a	Not detectable with current rules

Table 5.26: Rules for detecting untranslated single letter words

5.2.5.3 Results

The final algorithm was applied to detect and annotate untranslated words and incorrect acronyms in the 20,000 words of French-English MT output. Automatically flagged tokens and manual annotations were then compared. Figure 5.9 shows the number of errors detected and undetected, along with the number of false positives for each of the MT systems (Elliott et al., 2005).

Of the 203 manually annotated untranslated words and acronyms, 196 (96.5%) were automatically detected. In twelve of the sixteen sets of MT output, all annotated untranslated words were automatically detected, as shown in Table 5.27. The least successful results occurred in the case of the Reverso user manuals, where only five of the seven annotated errors were detected. All undetected errors involved the translation of

French acronyms into acceptable English words, or French words into acceptable English acronyms. These are shown in Table 5.27. In cases (1) and (2) errors in the MT output are not detectable using the methods described here. However, for cases (3) and (4), a rule was added to the algorithm to flag these tokens when found in lower case.

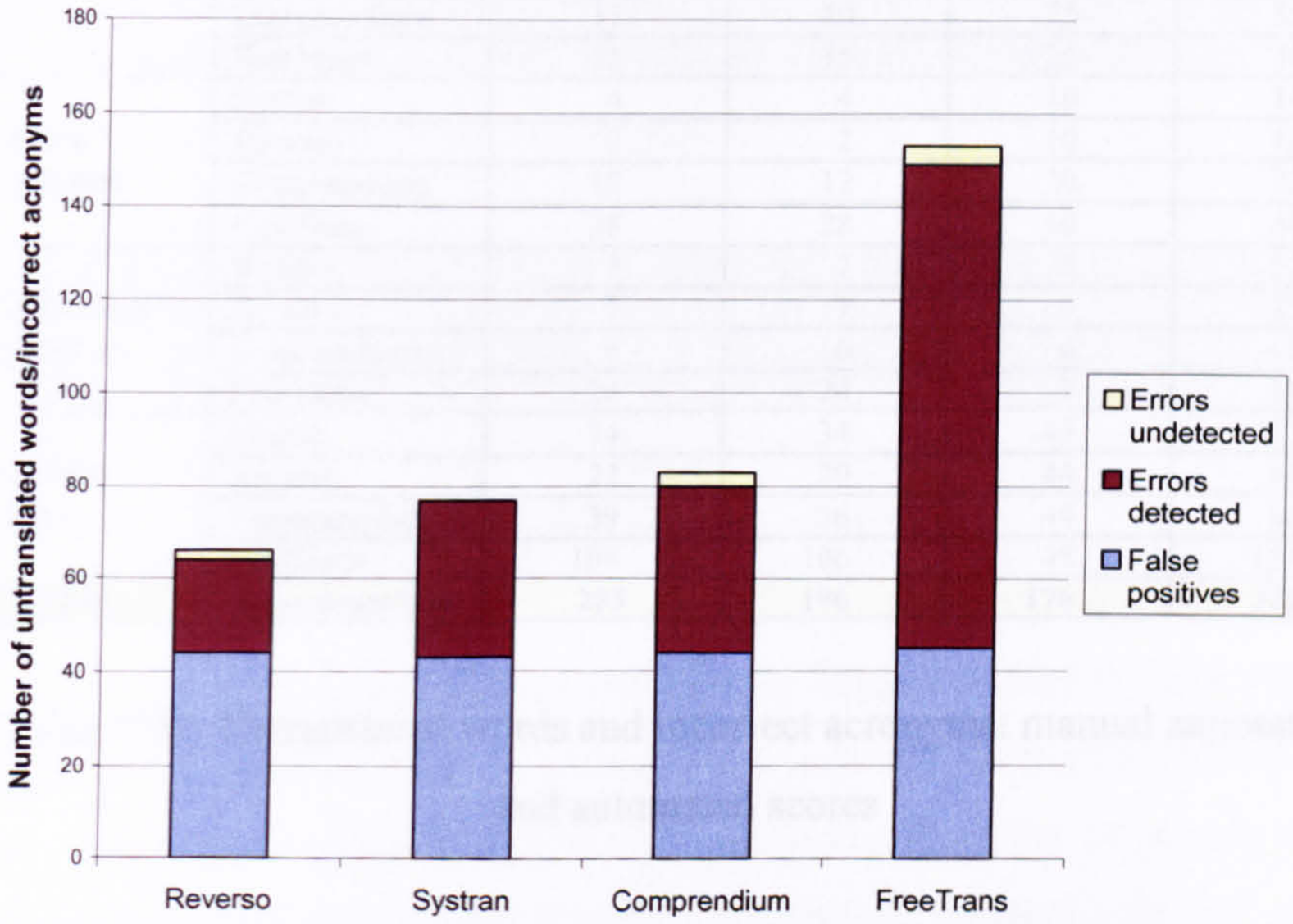


Figure 5.9: Untranslated words and acronyms: detection results by system

	French source	English MT output
1	AI (acronym)	HAVE
2	DOS (acronym)	BACK
3	ai (word)	ai (acronym)
4	un (word)	un (acronym)

Table 5.27: Undetected errors

Automated error counts were increased by a high number of false positives. Apart from one exception (a spelling error), all of these proved to be named entities (NEs) that did not appear in the English lexicon. For this reason, the number of false positives remains largely consistent in translations of the same source text by the four different systems. Table 5.28 shows results by system and text type (Elliott et al., 2005).

Text type	System	Manually annotated items	Number of annotated errors detected	Number of false positives	Auto score (no. of items detected)
User manuals	Systran	12	12	3	15
	Reverso	7	5	3	8
	Comprendium	10	8	3	11
	FreeTrans	41	39	3	42
FAQs	Systran	11	11	25	36
	Reverso	6	6	25	31
	Comprendium	11	10	25	35
	FreeTrans	15	15	25	40
Press releases	Systran	4	4	10	14
	Reverso	2	2	10	12
	Comprendium	12	12	10	22
	FreeTrans	28	28	10	38
Technical reports	Systran	7	7	5	12
	Reverso	7	7	6	13
	Comprendium	6	6	6	12
	FreeTrans	24	24	7	31
All text types	Systran	34	34	43	77
	Reverso	22	20	44	64
	Comprendium	39	36	44	80
	FreeTrans	108	106	45	151
TOTAL	**All systems/txts**	**203**	**196**	**176**	**372**

Table 5.28: Untranslated words and incorrect acronyms: manual annotations and automated scores

The consistency of the number of false positives across systems means that automated error counts still reflect the rank order of systems in terms of the number of manually annotated untranslated words and incorrect acronyms; this is also the case at the level of individual text types. However, the large number of named entities incorrectly flagged as 'untranslated' means that precision suffers greatly, as shown in Table 5.29 (Elliott et al., 2005). By adding the new NEs to the lexicon, precision increases to 100% for this sample. However, unseen data is always likely to contain new named entities not found in the lexicon.

System	Precision (including flagged NEs)	Precision (after adding NEs to lexicon)	Recall
Reverso	31.2	100	90.9
Systran	44.1	100	100.0
Comprendium	45.0	100	92.3
FreeTranslation	70.2	100	98.1

Table 5.29: Precision and recall by system

Precision can be improved by prompting the user to add named entities to the lexicon before running the program. This should not be so time-consuming, as findings from the sample data show that a large number of named entities are already included in the lexicon, and that new additions are most likely to be the names of specific products or people on which the texts focus. Even in texts containing many named entities, these are likely to be repeated several times; the 175 NEs incorrectly flagged as 'untranslated' in the whole sample involved only fourteen different tokens.

The introduction of named entity recognition software to filter out any NEs flagged as 'untranslated' was considered. However, a large number of correctly flagged untranslated words in the sample data were, in fact, named entities. Such a method would, therefore, improve precision, but at the considerable expense of recall.

In spite of low precision due to the large number of flagged named entities, results indicate that our automated identification of untranslated words and acronyms reflects human error annotations far more closely than Systran's automatic flagging of words not found in system dictionaries. Table 5.29 shows that for Systran our method achieved a precision of 44.1% and a recall of 100%. In comparison, Systran achieved only 14.6% for precision and 66.7% for recall based on the analyses discussed in 5.2.5.1.

5.2.5.4 Correlations with human judgements

Scores representing the number of automatically detected untranslated words and acronyms were compared with human judgements for fluency and adequacy (see Tables 4.3 and 4.4). Initial findings showed that the number of automatically detected untranslated tokens reflected the rank order of systems according to human adequacy judgements when all text types were taken into consideration. At the level of individual text types, a rank correlation was found only between automated scores and adequacy judgements for the user manuals.

Pearson's correlation co-efficient was subsequently computed to compare automated scores with human judgements. Two calculations are presented: Table 5.30 shows the value of r when comparing human judgements with (1) raw automated scores and (2) automated scores after the addition of named entities to the lexicon (Elliott et al., 2005).

Texts	Value of *r* : comparison with human fluency judgements		Value of *r* : comparison with human adequacy judgements	
	Raw automated scores	Scores after addition of NEs to lexicon	Raw automated scores	Scores after addition of NEs to lexicon
User manuals	-0.8628	-0.8628	-0.9972	-0.9972
FAQs	-0.9579	-0.9579	-0.9597	-0.9597
Press releases	-0.9310	-0.9310	-0.8747	-0.8747
Tech reports	-0.9422	-0.9264	-0.8618	-0.8388
All text types	**-0.9340**	**-0.9302**	**-0.9823**	**-0.9801**

Table 5.30: Correlations: untranslated words/acronyms and human judgements

Due to the consistent number of named entities found in translations of the same source texts by the different MT systems, the correlation coefficient remained the same for three of the text types, whether named entities were added to the lexicon or not. However, in the case of the technical reports, the correlation coefficient varied slightly due to differing numbers of named entities and other false positives in the MT output (see Table 5.28).

In the case of user manuals and FAQs, stronger correlations were found with human adequacy judgements, whereas for press releases and technical reports, stronger correlations were found with human fluency scores. Automated scores by system for the whole sample correlated more highly with human adequacy rather than fluency judgements.

A further experiment was conducted in an attempt to improve the correlations shown in Table 5.30. All single-letter untranslated words, listed in Table 5.26 and representing 28 of the 203 manual annotations, were removed from the algorithm, as all of these were determiners or pronouns attached to untranslated nouns or verbs and assumed to make little difference to human judgements of the MT output. The new set of automated scores was then compared with the human judgements. However, for eight of the ten samples, correlations were weaker than those shown in Table 5.30 offering no advantage over the initial algorithm.

5.2.5.5 Conclusions

An algorithm was developed to detect untranslated words and incorrect acronyms in French-English MT output by comparing words in the translations against an English lexicon and flagging all tokens not found. Further rules were added to the algorithm to detect untranslated single letter words and ordinal numbers.

This method identifies untranslated words and acronyms with access to the MT output alone and, therefore, in the same way as a native English speaker with no knowledge of the source language. For this reason, the approach should be extendible to other language pairs in which the target language is English: experiments with MT output from other source languages are described in Chapter 6. It should be noted that while this method does not detect some genuinely untranslated words, due to the fact that they exist as legitimate words in the target language (and were manually annotated as other error types), these could be identified as inappropriate by other error detection algorithms.

In total, 196 of the 203 manually annotated untranslated words and incorrect acronyms were detected using this method, equal to 96.5%. However, a large number of named entities were also flagged, due to the fact that they were not found in our English lexicon. This reflects our own human understanding and judgement, which is similarly reliant on updated knowledge for the correct interpretation of data. In spite of this, our automated error identification was much closer to human error annotations than Systran's flagging of words not found in system dictionaries.

In order to improve precision, named entities can be added to the lexicon before running the algorithm. However, with a sample of this size, Pearson's correlation coefficient showed that correlations between automated scores and human judgements remain largely unaffected, due the consistent number of false positives flagged in output from the different MT systems. Findings presented in 5.2.5.4 showed that scores based on automatically detected untranslated words and acronyms in the whole sample placed the four MT systems in rank order according to human adequacy judgements and also achieved a higher correlation coefficient with human adequacy rather than fluency scores.

5.3 Comparing correlations between automated scores and human judgements

Having explored correlations between automated scores for each individual error type and human judgements of fluency and adequacy, two further analyses were carried out. In the first, described in 5.3.1, correlations between automated scores for the five individual error types and human judgements are compared to determine which error type best predicts human scores. In the second analysis, presented in 5.3.2, scores based on the detection of combined error types are compared with human judgements to determine the set(s) of errors that best predict human scores based on this sample.

5.3.1 Correlations between automated scores for individual error types and human judgements

Correlation coefficients between automated scores for the five error types and human judgements were presented previously in this chapter: in Table 5.6 (outrageous words), Table 5.10 (inappropriate words), 5.14 (inappropriate prepositions), Table 5.23 (inappropriate pronouns) and Table 5.30 (untranslated words). These values were subsequently compared to determine which individual error type would best predict human judgements of fluency and adequacy across the different text types and for the sample as a whole. Figures 5.10 and 5.11 show correlation co-efficients for each text type, and for the whole sample, when comparing scores for individual error types with human fluency and adequacy judgements respectively. The strongest correlations are shown in the yellow highlighted range, the highest value being -1.

Findings based on this sample show that the number of automatically detected untranslated words best correlates with human judgements of both fluency and adequacy for three of the four text types and for the whole sample. Only in the case of technical reports do we find a slightly higher correlation between human scores and the number of detected inappropriate words. In contrast, the positive correlation with inappropriate prepositions in four of the five samples completely contradicts human judgements.

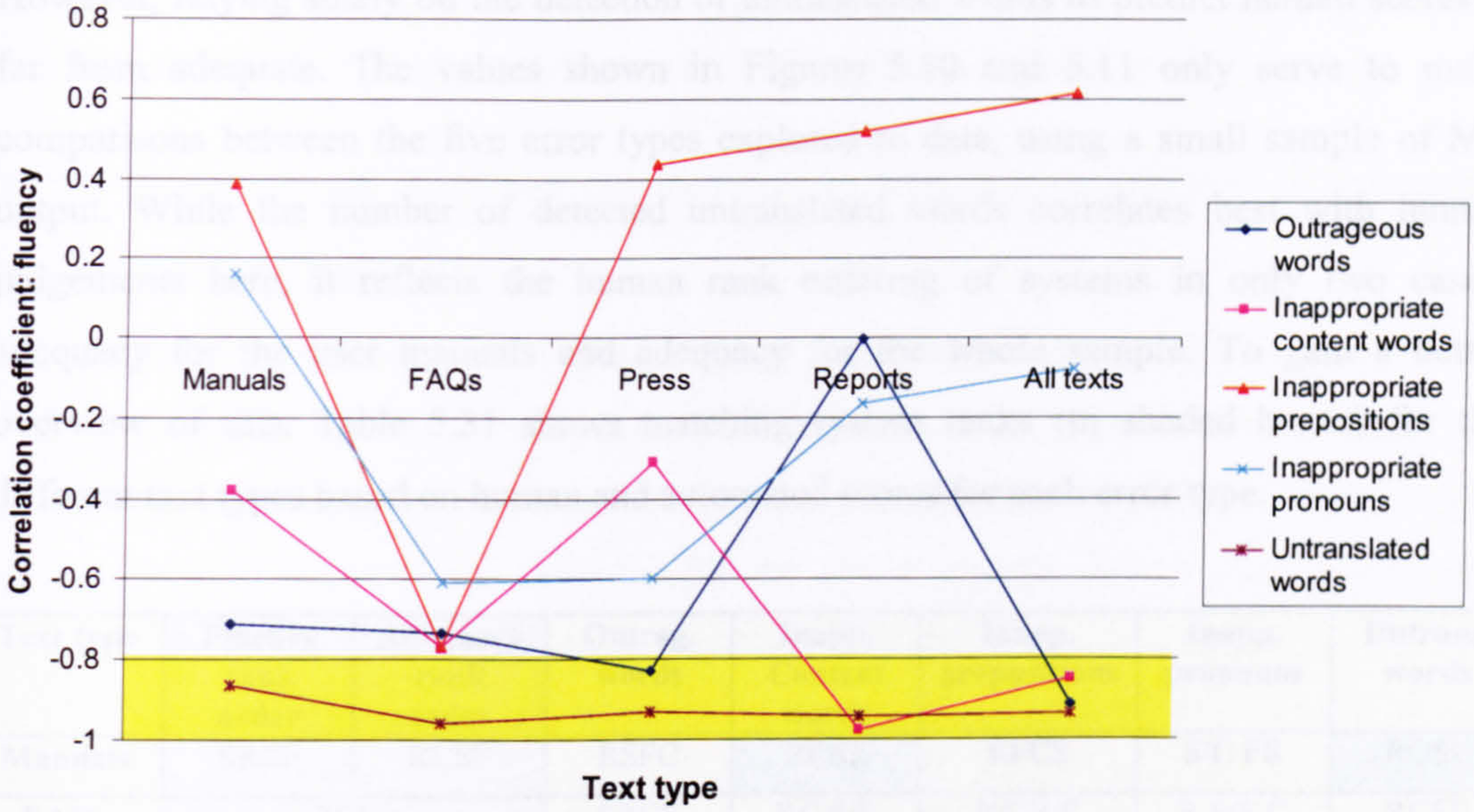


Figure 5.10: Correlations between human fluency judgements and automated scores for individual error types

Figure 5.11: Correlations between human adequacy judgements and automated scores for individual error types

However, relying solely on the detection of untranslated words to predict human scores is far from adequate. The values shown in Figures 5.10 and 5.11 only serve to make comparisons between the five error types explored to date, using a small sample of MT output. While the number of detected untranslated words correlates best with human judgements here, it reflects the human rank ordering of systems in only two cases: adequacy for the user manuals and adequacy for the whole sample. To gain a better overview of this, Table 5.31 shows matching system ranks (in shaded boxes) for the different text types based on human and automated scores for each error type.

Text type	Fluency rank order	Adequacy rank order	Outrag. words	Inapp. Content words	Inapp. prepositions	Inapp. pronouns	Untrans. words
Manuals	SRCF	RCSF	RSFC	RCSF	RFCS	S/C FR	RCSF
FAQs	RSCF		SRCF	RC S/F	RC S/F	R S/F C	RCSF
Press	SRCF		R S/C F	RSFC	FRSC	SC R/F	RSCF
Reports	SRCF		RCFS	SRCF	FSCR	SCFR	S/C RF
All texts	SRCF	RSCF	RSCF	RSCF	FRCS	SCFR	RSCF
S = Systran R = Reverso C = Comprendium F = FreeTranslation							
Corresponds with fluency ranking		Corresponds with adequacy ranking		Corresponds with fluency and adequacy ranking			

Table 5.31: Automated scores and human judgements: corresponding system ranks

Based on this sample, very few coinciding system ranks were found. The number of detected inappropriate content words would predict the rank order of systems for the user manuals and for the whole sample based on human adequacy scores and for the technical reports according to human fluency *and* adequacy judgements. Two other error types - outrageous words and untranslated words - would also predict system ranks in agreement with human judgements for adequacy when all texts are taken into account.

However, the automatic detection of one error type to predict human scores is not reliable: scores are insufficiently fine-grained, system improvements could significantly reduce the number of errors of a particular type, and one category of error cannot represent system

performance as a whole. Scores based on a combination of different error categories should be more reliable for comparing MT system performance in the long term.

5.3.2 Correlations between automated scores for combined error types and human judgements

Scores from the detection of combined error types were subsequently compared with human scores for fluency and adequacy to determine which set(s) of errors would best predict human judgements. Automated scores for each possible combination of error types were calculated for each text type and for the whole sample. Scores were unweighted, based purely on the frequency of each automatically detected error[5]. Pearson's correlation coefficient was then computed between each set of automated scores and human judgements. Error combinations representing the five strongest correlations with human scores for fluency and adequacy are shown in Figures 5.12 and 5.13 respectively. Note that the chart range in these figures corresponds to the yellow highlighted areas in Figures 5.10 and 5.11.

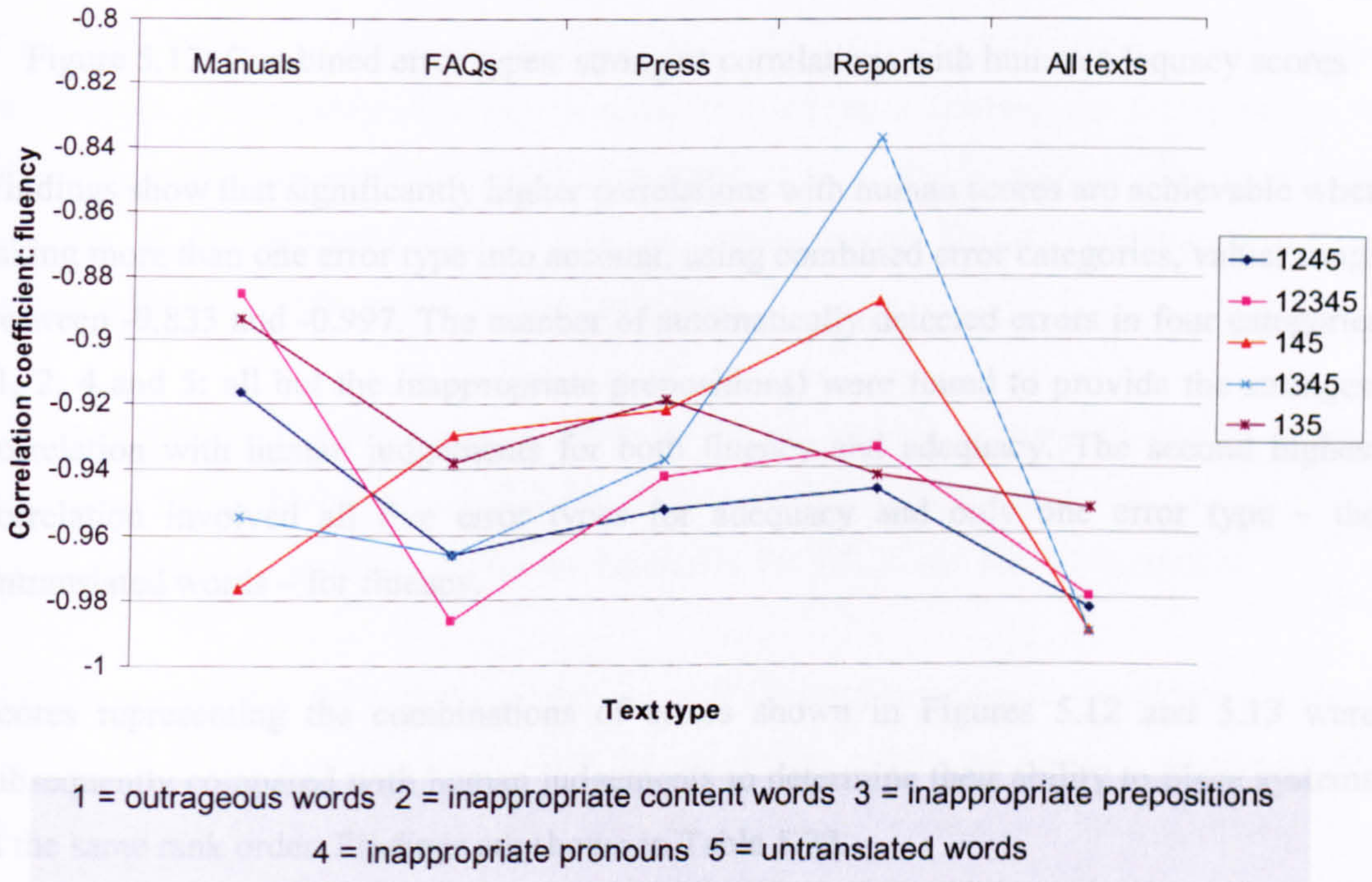
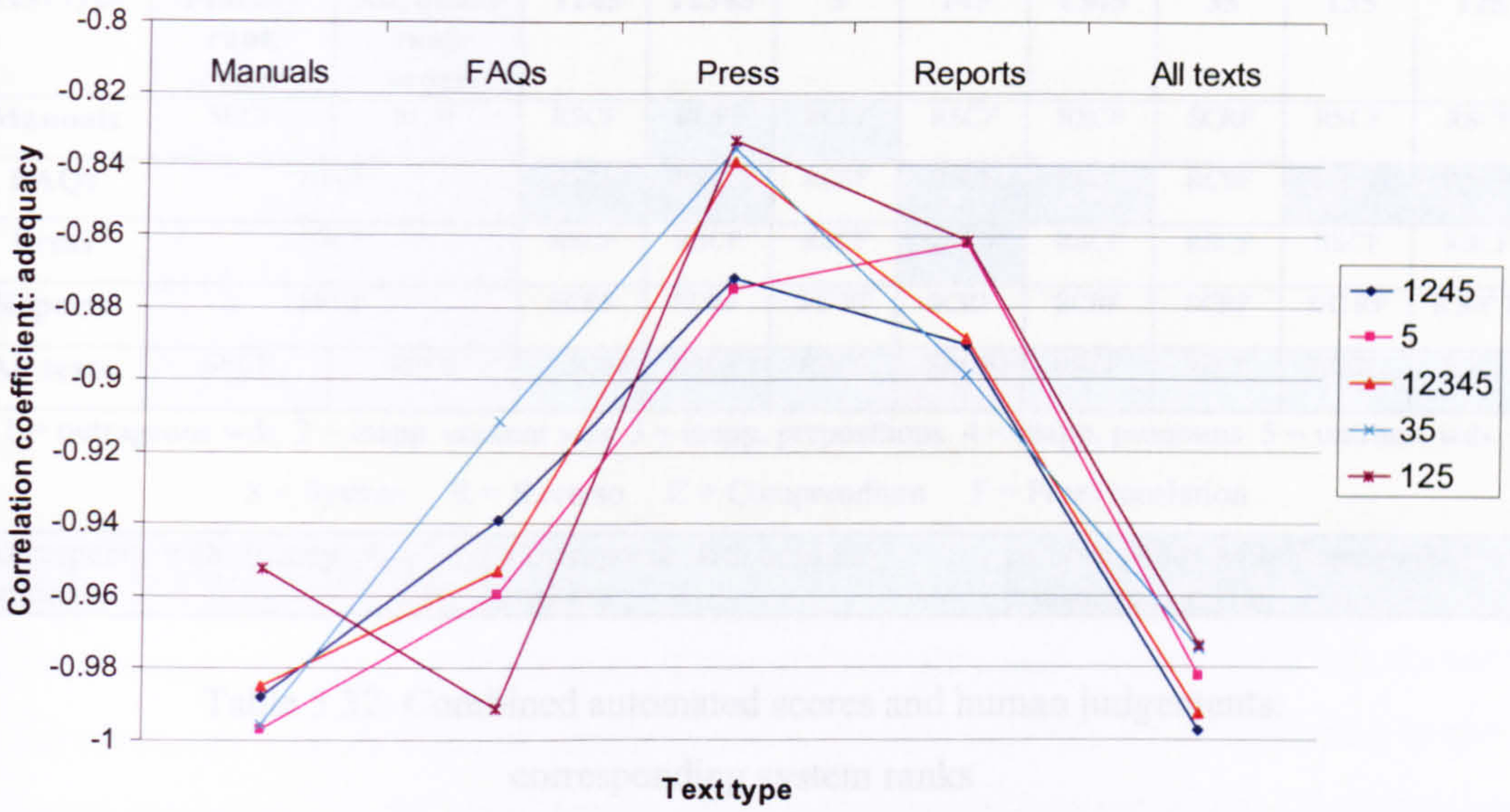


Figure 5.12: Combined error types: strongest correlations with human fluency scores

[5] Error weighting is a subject for further work.



1 = outrageous words 2 = inappropriate content words 3 = inappropriate prepositions
4 = inappropriate pronouns 5 = untranslated words

Figure 5.13: Combined error types: strongest correlations with human adequacy scores

Findings show that significantly higher correlations with human scores are achievable when taking more than one error type into account; using combined error categories, values range between -0.833 and -0.997. The number of automatically detected errors in four categories (1, 2, 4 and 5: all but the inappropriate prepositions) were found to provide the strongest correlation with human judgements for both fluency and adequacy. The second highest correlation involved all five error types for adequacy and only one error type – the untranslated words – for fluency.

Scores representing the combinations of errors shown in Figures 5.12 and 5.13 were subsequently compared with human judgements to determine their ability to place systems in the same rank order. Findings are shown in Table 5.32.

Text type	Fluency rank order	Adequacy rank order	1245	12345	5	145	1345	35	135	125
Manuals	SRCF	RCSF	RSCF	RCSF	RCSF	RSCF	RSCF	SCRF	RSCF	RSCF
FAQs	RSCF		RSCF	RSCF	RCSF	RSCF	RSCF	RCSF	RSCF	RSCF
Press	SRCF		RSCF	RSCF	RSCF	S/R CF	RSCF	RSCF	RSCF	RSCF
Reports	SRCF		SCRF	SCRF	S/C RF	SCRF	SCRF	SCRF	S/C RF	R S/C F
All texts	SRCF	RSCF	RSCF	RSCF	RSCF	SRCF	RSCF	RSCF	RSCF	RSCF
1 = outrageous wds 2 = inapp. content wds 3 = inapp. prepositions 4 = inapp. pronouns 5 = untrans. wds S = Systran R = Reverso C = Comprendium F = FreeTranslation										
Corresponds with fluency ranking		Corresponds with adequacy ranking				Corresponds with fluency and adequacy ranking				

Table 5.32: Combined automated scores and human judgements: corresponding system ranks

The shaded boxes, representing corresponding ranks in Table 5.32, show that all of the above error combinations are able to predict the rank order of competing systems based on the whole sample; in seven out of eight cases, these error combinations predict human judgements of adequacy. Bearing in mind that we have focused on the detection of semantic rather than syntactic errors, this result is very encouraging. It also proves the potential of an automated method that can predict human fidelity judgements without access to the source text or to a human translation.

The most successful combination of errors for system ranking involves all five categories: automated scores based on the five error types predict system ranks that coincide with human adequacy judgements for the user manuals, the FAQs and the whole sample.

Stronger correlations with human scores (of either fluency or adequacy) could be achieved by working to improve the current error detection rules and by adding new algorithms for the detection of other error types. However, we must bear in mind that the algorithms have been tested on a small data sample, due to the fact that all texts needed to be evaluated by humans and manually annotated with errors, which is very costly and time-consuming. Stronger correlations with human judgements or, at least, better predictions of the human ranking of MT systems across all of the text types, may be achievable with a larger sample.

5.4 Automated error detection: conclusions

We have seen in Chapter 3 that the human evaluation of MT output is expensive and time-consuming and that the most widely used automated evaluation method – BLEU (Papineni et al., 2001) and its various adaptations – requires typically four human translations of every text to be evaluated. In response, this chapter has presented an investigation of the potential of a method to automatically evaluate MT output from competing systems without access to the source text or to any human translation. Research was carried out to establish whether it would be possible to reliably detect errors automatically in MT output based solely on the target text and to determine any correlation between the frequency of these detected errors and human judgements of fluency and adequacy. Initial work has focused on the automatic identification of semantic errors, divided into five types according to their detection methods: 'outrageous' words, other inappropriate content words, inappropriate prepositions, inappropriate pronouns and untranslated words and acronyms.

Algorithms were developed and tested on machine translations of 12 texts of around 400 words by each of four competing systems, providing a total of approximately 20,000 words. MT output of four text types was randomly selected from a corpus designed specifically for this research. Each machine translation was manually annotated with error types based on an error classification scheme designed for this purpose and had been evaluated by humans to provide fluency and adequacy scores at the segment level. This would enable (1) the accuracy of the automated error detection methods to be tested by comparing automated and manual annotations and (2) the comparison of the number of automatically detected errors and human scores to explore the possibility of automatically predicting human judgements.

Algorithms for the detection of each error type were developed after detailed manual analyses of the MT output and, in all cases, the use of other much larger corpora from which word lists were automatically extracted or lists of rules manually compiled. These corpora include millions of words of original English texts and MT output from different systems. A list of corpora, lexicons and error lists developed for this work is provided in Appendix 7.

5.4.1 Summary of error detection results

Results showed that a significant number of errors are detectable using rule-based automated methods, when the source text or a human translation is not available. The number of automatically detected errors of each type is shown in Figure 5.14.

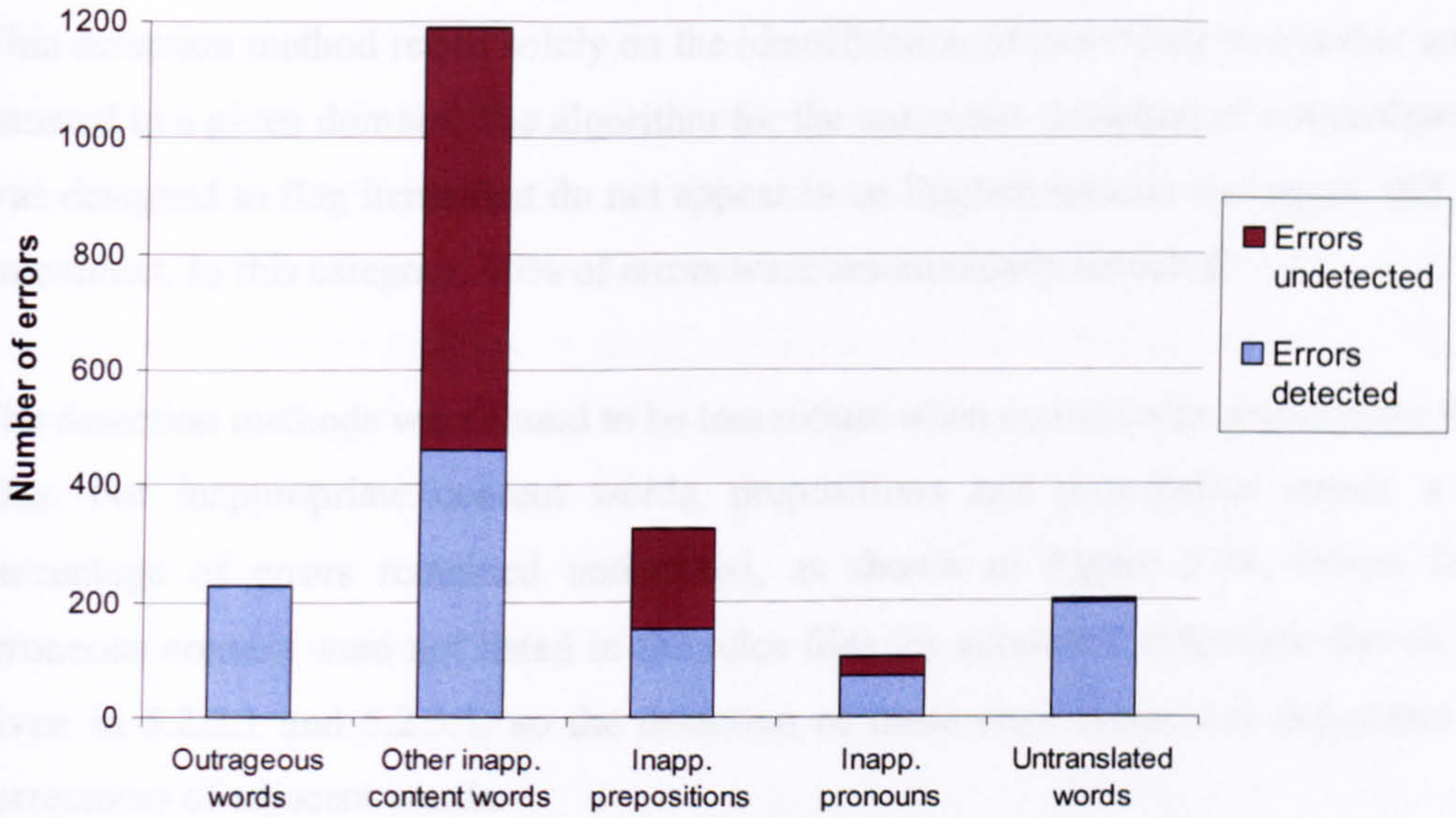


Figure 5.14: Automatically detected errors compared with manual annotations

Of the 2,049 manually annotated errors in the five categories for which detection methods were developed, 1,110 (54%) were automatically detected without the requirement of a human translation or access to the source text. This amounts to 24% of all annotated errors in the sample and 0.85 errors per segment. In the case of outrageous words, inappropriate content words and inappropriate prepositions, no correct items were automatically flagged as erroneous. However, for pronouns, eight false positives were detected in addition to the 72 correctly identified errors and in the case of untranslated words, named entities were incorrectly flagged when they were not found in the English lexicon. The number of NEs detected as untranslated words was consistent in translations of the same source texts by the different systems, so automated scores could still reflect comparative error frequency by system, whether named entities were detected or not. However, the user could be prompted

to add NEs to the lexicon before running the program to prevent items from being incorrectly flagged as untranslated.

During the course of this research, it has become clear that a greater percentage of manually annotated errors are automatically detectable when the algorithms do not rely on context. For example, all 52 words that had been manually annotated as 'outrageous' were detected (along with a further 178 words which had also been manually annotated as inappropriate). This detection method relied solely on the identification of individual words that are highly unusual in a given domain. The algorithm for the automatic detection of untranslated words was designed to flag items that do not appear in an English lexicon and again, did not rely on context. In this category, 96% of errors were automatically detected.

The detection methods were found to be less robust when context was required for the rules files. For inappropriate content words, prepositions and pronominal errors, a greater percentage of errors remained undetected, as shown in Figure 5.14. Errors found in erroneous context were not listed in the rules files for automatic detection, due to reasons given in 5.2.2.1 and 5.2.3.1, so the detection of these error types was dependent on the correctness of adjacent words.

In addition to successfully detecting a significant number of errors, it was also found that for four out of the five error types, error corrections could be automatically inserted into the translations if desired. This is possible when the detection method relies on a list of erroneous word combinations to identify in a given sample. It is not possible to incorporate this feature in the detection method used for the majority of untranslated words. This is due to the fact that these errors are identified when tokens are *not* found in a word list.

5.4.2 Summary of correlations with human scores and ranks

Having devised automated detection methods for five error types, correlations were investigated between unweighted error frequencies and human quality judgements. For individual error types, the strongest (negative) correlation was found between the number of automatically detected untranslated words and human scores for both fluency and adequacy. This can be seen in Figures 5.10 and 5.11. Additionally, based on the whole

sample, the frequency of automatically detected outrageous words, other inappropriate content words or untranslated words would place translations by the four MT systems in the correct rank order according to human adequacy judgements. See Table 5.31.

Correlations between human and automated scores were found to improve when combining the frequencies of different error types. The highest correlation co-efficient was found between human fluency *and* adequacy scores and the frequency of four error types (all apart from the inappropriate prepositions). This is shown in Figures 5.12 and 5.13. However, in terms of placing the sets of translations in the correct rank order in accordance with human scores, the frequency of all five error types should be computed. This would place the four systems in rank order according to human adequacy judgements, based on the whole sample; in addition, it would correctly rank two sub-sets from the sample: the user manuals (in agreement with human adequacy scores) and the FAQs (according to fluency *and* adequacy). See Table 5.32.

5.4.3 Summary of findings

Work described in this chapter has shown that errors in MT output can be automatically detected without the source text or human translations. Furthermore, it is possible when detecting some error types to extend the algorithms to offer error corrections. Algorithms developed to date are able to generate scores, based on error frequency, that can predict the human ranking of four MT systems when a sample of twelve texts in the same domain is evaluated.

Correlations with human scores were found to improve as more error types were detected; further work will involve the development of methods to automatically detect other error types with the goal of improving these correlations and producing more finely differentiated scores, rather than just ranks, based on output from competing systems of varying levels of quality. Suggestions for future work are described in Chapter 7.

The current approaches should be portable to other language pairs in which the target language is English; an investigation of the ability of the algorithms to detect errors in MT output from other source languages is presented in Chapter 6.

Chapter 6

Extending methods to other language pairs

As stated in 5.1.1, a major advantage of an automated evaluation method that relies solely on the analysis of the MT output is that it should be easily extendible to other language pairs in which the target language is the same. This chapter presents findings obtained when existing algorithms for the detection of the five error types were tested on English output translated from three other source languages: German, Italian and Spanish. Tests were carried out to determine the extent to which the algorithms would need to be adapted to detect a similar proportion of errors in English MT translated from other languages. Investigations of correlations with human scores were not essential at this point, as the primary concern was to establish whether the algorithms would identify any errors at all and if so, to discover the level of error similarity in translations from different source languages.

Twelve original texts (or extracts from longer texts) in each of the three source languages were selected from the web to produce three samples, comparable in terms of size, text type and domain to the Fr-En computing corpus (see 4.4). These were translated into English by two available systems: Systran Premium 5.0 and FreeTranslation, representing one of the best performing systems and the worst performing system respectively in our human and automated evaluations (see 4.4 and Chapter 5) for the French-English pair. An initial analysis of translations from German, Italian and Spanish showed that the quality of output from Systran remained higher, although some sentences in the output from FreeTranslation (for all three language pairs) were superior in terms of syntax and word choice. The example of the German-English corpus is shown in Figure 6.1.

Without access to the source texts, the 10,000 words of English MT output in each of the three samples were manually annotated with the five error types: outrageous words, other inappropriate content words, inappropriate prepositions, inappropriate pronouns and untranslated words/incorrect acronyms. The existing error detection algorithms were then

used, without any adaptation, to automatically identify errors in the three samples. Automated and human annotations were subsequently compared.

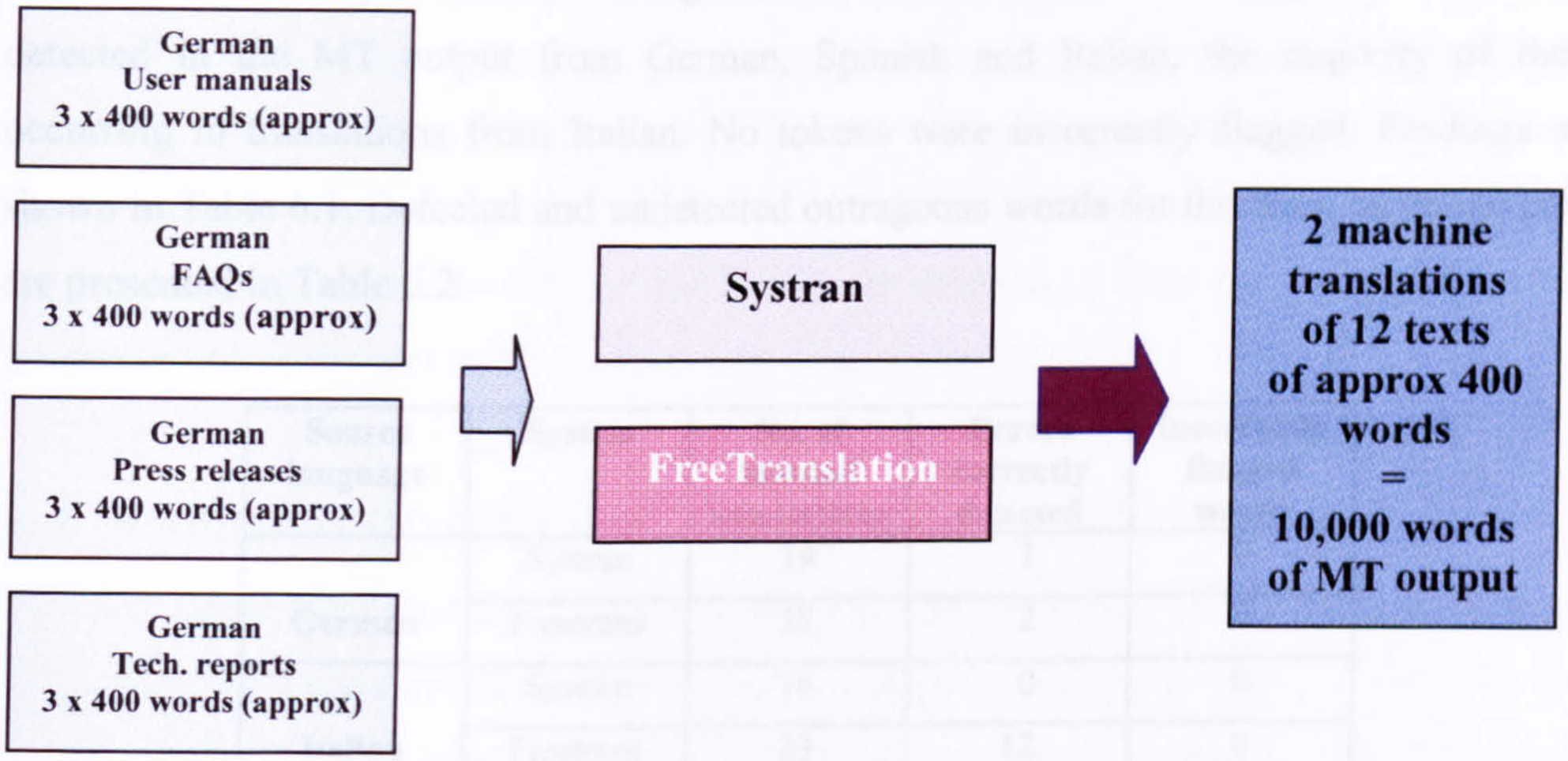


Figure 6.1: German-English corpus for MT evaluation

Findings, observations and suggestions for improving the existing rules to cater for additional language pairs are presented for each of the five automated error detection methods. The detection of outrageous words is discussed in 6.1, other inappropriate content words in 6.2, inappropriate prepositions in 6.3, pronouns in 6.4 and untranslated words and acronyms in 6.5. Conclusions are presented in 6.6.

6.1 Automatic detection of outrageous words

The MT output from German, Italian and Spanish was automatically compared against the outrageous lexicon developed to identify mistranslations from French into English. This list had been generated by automatically comparing a 750,000 word sample of French-English MT output in the computing domain with a three million word corpus of comparable original English texts. Legitimate English words found in the MT output but not in the original English texts were automatically extracted to create a list of candidate 'outrageous' words, whose appearance is highly unusual in a given text type and domain. The method is described fully in 5.2.1.4. Following manual checks to remove unsuitable items and insert, for example, additional verb conjugations, the lexicon contained 390 words. For the

French-English language pair, all tokens manually annotated as 'outrageous' were detected using this method.

With the list developed for French-English MT, a small number of outrageous words were detected in the MT output from German, Spanish and Italian, the majority of these occurring in translations from Italian. No tokens were incorrectly flagged. Findings are shown in Table 6.1. Detected and undetected outrageous words for the three language pairs are presented in Table 6.2.

Source language	System	No. of manual annotations	Errors correctly detected	Incorrectly flagged words
German	Systran	19	1	0
	Freetrans	18	2	0
Italian	Systran	16	0	0
	Freetrans	23	12	0
Spanish	Systran	7	0	0
	Freetrans	10	1	0

Table 6.1: Number of automatically detected outrageous words in output from German, Italian and Spanish

Language pair	Detected outrageous words	Undetected outrageous words
German-English	seized, sink	haven, warden, footstep, flared, ventilator, closet, traitorous, treacherous, traitor, chopped, robbery, creeping, singe, contestant, slinking, paltry, silos, baking, blazes, bogs, stalemates
Italian-English	glue, throw, tongue(s)	patrimony, commando, pigeonhole, rifle, fairies, zipper, congenital, forehead, paddles, dumb, solariums, trickiness
Spanish-English	baptized	booth, servant(s)

Table 6.2: Detected and undetected outrageous words in output from German, Italian and Spanish

The examples in Table 6.2 show that quite different outrageous words occur in English MT output depending on the source language and that in order to automatically detect a

significant number of these errors found in translations from other source languages (a) the current word list, based on English MT from French, would have to be extended or (b) a separate word list would need to be created for each language pair. This would involve the same procedure used to compile the French-English outrageous lexicon, whereby large corpora of machine translations from the relevant source languages would be automatically compared with original English texts in the same domain to identify candidate mistranslations. Methods are described in 5.2.1.4. This process is mainly automated; only the final stage involves the manual verification or deletion of entries in the automatically generated outrageous lexicon.

6.2 Automatic detection of other inappropriate content words

The MT output from German, Italian and Spanish was automatically compared against the list containing inappropriate content words developed for French-English. This list had been compiled by manually analysing a 90,000 word sample of French-English MT output from three available systems (see 5.2.2.1) and selecting inappropriate collocations for inclusion. The updateable list contained 1490 items. Examples are provided in 5.2.2. For the French-English language pair, 38% of the inappropriate content words that had not been flagged as 'outrageous' were detected using this method.

Using the French-English inappropriate word list, a relatively small number of errors were detected in the MT output from German, Italian and Spanish. Just as in the case of outrageous words, the majority of these were detected in translations from Italian. No tokens were incorrectly flagged. Manual and automated annotations are shown in Table 6.3. Detected inappropriate content words for the three language pairs can be seen in Table 6.4.

As with the outrageous words, findings show that while a small number of errors are detectable using the French-English error file, the majority are not. This means that in order to achieve results comparable to the French-English language pair, the current inappropriate word list would have to be extended to include errors found in MT from other source languages or a separate list would need to be compiled for each language pair.

Source language	System	No. of manual annotations	Errors correctly detected	Incorrectly flagged words
German	Systran	299	5	0
	Freetrans	410	2	0
Italian	Systran	459	12	0
	Freetrans	531	20	0
Spanish	Systran	299	10	0
	Freetrans	297	10	0

Table 6.3: Number of automatically detected inappropriate content words in output from German, Italian and Spanish

Output from German	Output from Italian	Output from Spanish
record the file (save)	record the file (save)	eliminated the file (deleted)
on the apparatus (telephone)	eliminated the file (deleted)	eliminate the sentence (delete)
the spreading (distribution)	register the document (save)	register the file (save)
attain the menu (reach)	in shape of (in the form of)	in shape of (in the form of)
	verified that (checked)	
	I would want (like)	
	is very spread (widespread)	

Table 6.4: Detected inappropriate content words in output from German, Italian and Spanish

6.3 Automatic detection of inappropriate prepositions

The list of errors used to detect inappropriate prepositions was compiled during the manual analysis of a 90,000 word sample of French-English MT output from three available systems (see 5.2.3.1) and contained 572 entries. Examples are provided in 5.2.3. When automatically comparing the list with the Fr-En computing corpus, 47% of prepositional errors were detected. However, when used to identify inappropriate prepositions in output translated from German, Italian and Spanish, very few errors were automatically detected. No errors were identified in the output from German. Manual and automated annotations are shown in Table 6.5. The detected inappropriate prepositions in output from Italian and Spanish involved only three error types, as shown in Table 6.6.

Source language	System	No. of manual annotations	Errors correctly detected	Incorrectly flagged words
German	Systran	65	0	0
	Freetrans	90	0	0
Italian	Systran	70	0	0
	Freetrans	61	1	0
Spanish	Systran	83	3	0
	Freetrans	83	4	0

Table 6.5: Number of automatically detected inappropriate prepositions in output from German, Italian and Spanish

Output from Italian	Output from Spanish
chosen in (from)	in (on) this occasion of (in) Windows

Table 6.6: Detected inappropriate content words in output from German, Italian and Spanish

Of the five error detection methods developed, this one proved to be the least successful when tested on output translated from other languages. Machine translations from German, Italian and Spanish would have to be analysed to create lists of inappropriate prepositions for each language pair.

6.4 Automatic detection of inappropriate pronouns

As discussed in 5.2.4.2, a number of inappropriate pronouns found in MT output from French were clearly due to multiple translation possibilities for third person pronouns: the attribution of gender to inanimate objects in French sometimes caused 'il' or 'elle', for example, to incorrectly translate into 'he' or 'she' rather than the required 'it'.

In Italian and Spanish, on the other hand, subject pronouns are normally omitted, so when verbs are used in the third person, an MT system must insert a subject pronoun, which

could legitimately be 'he', 'she' 'it' or 'you'. Furthermore, just as in French, singular object pronouns in both languages can have multiple legitimate translations: 'him', 'her' and 'it'. Similar problems arise when translating from German into English. The subject pronoun 'Sie' (upper case 'S'), for example, translates into 'you', whereas 'sie' translates into 'they'. However, their respective verb conjugations are the same. For all of the above reasons, it was anticipated that the method used to automatically flag inappropriate pronouns in English output from French (see 5.2.4.2) would also detect a number of errors in output from German, Italian and Spanish. For the French-English language pair, 72/104 (69%) of inappropriate pronouns were detected using this method, and eight items were incorrectly flagged.

The inappropriate pronoun list developed for French-English, containing 23 entries, was automatically compared against the MT output from German, Italian and Spanish and all matches flagged. Manual annotations and automatically identified errors are shown in Table 6.7. Detected and undetected inappropriate pronouns in output from each source language are shown in Table 6.8.

Using the method originally developed to detect incorrect pronouns in French-English MT, 21% of errors were detected in output from German, 29% in output from Italian and 39% from Spanish. However, five items were incorrectly flagged in the German FAQs and Spanish technical reports.

Source language	System	No. of manual annotations	Errors correctly detected	Incorrectly flagged words
German	Systran	41	7	1 (his)
	Freetrans	43	11	1 (he)
Italian	Systran	18	10	0
	Freetrans	20	1	0
Spanish	Systran	34	14	2 (him/his)
	Freetrans	17	6	1 (him)

Table 6.7: Number of automatically detected inappropriate pronouns in output from German, Italian and Spanish

Detected errors	Examples of undetected errors	Correct translation
Output from German		
he himself , what 'one' as subject with verbs on error list	, those / , the / , its them / they your its 'one' as subject pronoun with verbs not on error list	, which you its your
Output from Italian		
he she him , what	them / they It that you It That	you What they This It
Output from Spanish		
he she him his , what	, that / who / whom it itself that its	which you you / yourself what their

Table 6.8: Detected and undetected inappropriate pronouns in output from German, Italian and Spanish

Findings suggest that the pronominal error list used for French-English could also be adopted for the detection of errors in output from German, Italian and Spanish. However, further entries should be added to the list where possible, to detect additional errors shown in Table 6.8, which were not found in output from French.

6.5 Automatic detection of untranslated words and acronyms

For the purpose of this work, in which error detection does not rely on the source text or any human translation, untranslated items are defined as words or acronyms which are not legitimate in the English language and are, therefore, detectable by examining the MT output in isolation. For the French-English language pair the MT output was automatically compared, after pre-processing to remove non-words[1], against an English lexicon to flag all tokens not found. This method was selected due to its anticipated extendibility to other language pairs in which the target language is English and was designed to identify

[1] For the purpose of this thesis, non-words refer to any character string containing one or more non-alphabetic characters (excluding punctuation). These include email addresses, URLs and file names.

untranslated words in much the same way as a native English speaker who has no knowledge of the source language. The algorithm is described fully in 5.2.5.2.

The method for capturing untranslated words and acronyms was used to compare the MT output from German, Italian and Spanish against the English lexicon, described in 5.2.1.1, and to automatically flag all tokens not found. Annotated tokens in the returned text were then compared with the manually annotated untranslated words. Findings are summarised in Table 6.9.

Source language	System	No. of manual annotations	Errors correctly detected	Errors undetected	Automated score (total number of tokens detected)	Incorrectly flagged tokens and spelling errors
German	Systran	70	50	20	115	65
	Freetrans	80	69	11	134	65
Italian	Systran	83	75	8	157	82
	Freetrans	76	69	7	141	72
Spanish	Systran	116	87	29	162	75
	Freetrans	122	86	36	146	60

Table 6.9: Number of automatically detected untranslated tokens in output from German, Italian and Spanish

6.5.1 Explanations for non-detection of untranslated words and incorrect acronyms

The manually annotated untranslated words and incorrect acronyms that were not detected using the existing method can be divided into three types. These are described below and shown in Table 6.10.

1. A large number of untranslated words and acronyms were not detected due to the presence of German, Italian and Spanish words in the English lexicon. This problem is solvable by removing these tokens from the lexicon, although due to its size (over 320,000 entries) this will not be trivial.

2. Three items were not detected because they were excluded from analysis as a result of tokenisation and pre-processing to remove 'non-words'. These were 17de in Spanish (17th), e/o in Italian (and/or) and Z.B in German (E.G). As with the French-English language pair, rules need to be added to enable the automatic recognition of untranslated ordinal numbers and other items that are not detected by simply comparing tokens in the MT output against the English lexicon.

3. A small number of items cannot be detected using the current method. These are (a) incorrect acronyms, whether untranslated or mistranslated, that form acceptable words or acronyms in the English language and (b) acronyms containing full stops as separators (such as O.A.M.I found in output from Spanish) which are omitted from analysis during pre-processing (see 5.2.5.2). However, the use of full stops in acronyms is very rare; this was the only case found in all of the MT output evaluated for this research.

Language	Words/Acronyms to be removed from lexicon	Tokens requiring additional rules	Tokens undetectable using this method	No. of undetectable tokens in sample
German	Korpus industrielle Nomen	Z.B	QA FLEW CTO	18
Italian	crea tal stampa milano lingua yagi di ecc verrai usi ultima	e/o	0	0
Spanish	de INFORMACION al EDUCACION por dia ter	17de	O.A.M.I YOU OEP	8

Table 6.10: Undetected untranslated words and incorrect acronyms in output from German, Italian and Spanish

By removing non-English words from the lexicon and adding rules to enable the detection of the tokens described in (2) above, it is possible to detect 88% of untranslated words and incorrect acronyms in the sample of MT output from German, 100% of errors translated from Italian and 96% of errors in the output from Spanish using the current method. These

results are comparable to those obtained for French-English, where 96.5% of untranslated tokens were automatically flagged.

6.5.2 Explanations for incorrectly flagged tokens

When analysing output from German, Italian and Spanish, four reasons were identified for the incorrect flagging of tokens as 'untranslated', as shown below. Each of these is discussed in turn.

1. Named entities not found in the English lexicon
2. Legitimate words not found in the English lexicon
3. Spelling errors or 'invented' words caused by the MT system
4. Erroneous concatenation of words by the MT system

6.5.2.1 Incorrect flagging of named entities

As in the case of output from French, a large number of NEs were incorrectly flagged as 'untranslated' in MT output from German, Italian and Spanish, as they did not appear in the English lexicon. In spite of this problem, the number of NEs found in translations of the same source texts by each system was consistent, so this did not prevent the correct ranking of systems in terms of the frequency of this error type. Frequencies of NEs are shown by text type in Table 6.11.

Text type	Output from German		Output from Italian		Output from Spanish	
	Systran	FreeTrans	Systran	FreeTrans	Systran	FreeTrans
Manuals	31	31	12	12	48	48
FAQs	3	2	25	23	5	5
Press	19	19	17	17	1	1
Reports	5	5	7	7	0	0
Total	**58**	**57**	**61**	**59**	**54**	**54**

Table 6.11: Number of named entities incorrectly flagged as 'untranslated' in output from German, Italian and Spanish

In order to prevent this problem, NEs should be added to the English lexicon during the evaluation. The most efficient way to do this is when the unique list of 'untranslated' tokens is generated for a given sample, but before tokens from the list are flagged in the MT output (ie. between steps 5 and 6 in the algorithm described in 5.2.5.2). For the 10,000 words of MT output from German, Italian and Spanish, the unique lists of untranslated tokens contained on average 134 entries per language pair. NEs can be deleted from the list so that they are not flagged as errors, and inserted into the English lexicon at the same time. Due to a large amount of repetition, the number of NEs that would need to be added to the lexicon is relatively small. In this sample there were 16 different incorrectly flagged NEs in output from German, 16 in output from Italian and 9 from Spanish. (NB. The additional NEs found in output from Systran are a result of the system's tendency to occasionally repeat phrases.)

6.5.2.2 Legitimate words not found in the English lexicon

In total, 33 legitimate English words were flagged as 'untranslated' in output from the three source languages due to the fact that they did not appear in the English lexicon. Examples are *renouncement, bulkheading, communiqué, trustworthily, standardisations, plaintext, portlet, spidering, logons, codecision* and American English spellings such as: *coorganized, initializes, memorization.* Steps should be taken to ensure that these tokens, including American-English spellings, are added to the lexicon to prevent their detection as 'untranslated' words.

6.5.2.3 Flagging of spelling errors and invented words

There were 18 instances of spelling errors or invented words in the output from the three different source languages. Most of these were found in output from Systran, as shown in Table 6.12. Although these tokens are not untranslated, their detection as spelling errors is useful, as it indicates how improvements can be made to system dictionaries, whether by developers or users. The method developed for the detection of untranslated words will also enable the detection of errors in this additional group.

System	Output from German	Output from Italian	Output from Spanish
Systran	processings	adressees oing immensest usefullness furnitures dissimation certifyd	attachs x 2 nonwished characteristicses unskillful
FreeTrans	processings trustworthyer	---	attachs x 2 annexs

Table 6.12: Spelling errors and invented words in output from German, Italian and Spanish

6.5.2.4 Flagging of erroneous word concatenation

Only one instance was found in which Systran had generated two adjacent words without inserting a space. As a result, the token was not found in the English lexicon. Although this may appear to be an isolated occurrence, the author has seen many such instances in other samples of output from Systran and it is useful for post-editors, for example, to identify these cases so that corrections can be made.

6.5.3 Summary of required improvements

Based on the above observations, it is clear that the accuracy of the method for the automatic detection of untranslated words and incorrect acronyms can be improved by:

1) removing non-English words from the lexicon;
2) writing additional rules to detect errors which are not flagged due to tokenisation and pre-processing;
3) adding NEs to the lexicon to prevent incorrect flagging;
4) adding legitimate words to the English lexicon when they are discovered.

Detection results for untranslated tokens before and after these improvements were made, based on the current sample, are compared in Table 6.13. Findings show that the algorithm for detecting untranslated words is the most successful of the five error detections methods when applied to other language pairs, in that it enables the detection of the highest percentage of errors in output from other source languages. NB. Precision is based on the

detection of untranslated words and acronyms; this would be 100% if we accept the detection of spelling errors and invented words using the same algorithm.

Language pair	Results before improvements		Results after improvements	
	Precision	Recall	Precision	Recall
German-English	47.8%	79.3%	97.1%	88.0%
Italian-English	48.3%	90.6%	95.8%	100%
Spanish-English	56.2%	72.7%	96.6%	96.6%

Table 6.13: Detection results before and after improvements for untranslated words and incorrect acronyms

6.6 Conclusions

After developing methods for the automated detection of five error types in English MT output from French, the same algorithms were tested on output translated from three other source languages: German, Italian and Spanish. These experiments were carried out to identify the extent to which the algorithms would need to be adapted to detect a similar proportion of errors in MT output translated from other languages. Detection results for all four language pairs are summarised in Table 6.14. (Results for untranslated words are shown before changes are made to the English lexicon.)

	Error type	German-English	Italian-English	Spanish-English	French-English
1	Outrageous words	8.1%	30.8%	5.9%	100%
2	Inappropriate content words	0.9%	3.2%	3.3%	38.6%
3	Inappropriate prepositions	0%	0.8%	4.2%	47.4%
4	Inappropriate pronouns	21.4%	28.9%	39.2%	69.2%
5	Untranslated words/acronyms	79.3%	90.6%	72.7%	96.5%

Table 6.14: Percentage of each error type detected for all language pairs

Findings have shown that for the detection of the first three error types in Table 6.14, the existing error files compiled for French-English are not adequate for the detection of errors

in output from German, Italian or Spanish. In order to extend the current methods, a new error list would need to be compiled for each language pair, even though the target language is the same.

The method for the detection of inappropriate pronouns was more successful, largely due to the fact that German, Italian and Spanish, in addition to French, are gender-based languages. This means that some of the same pronominal errors occur in output from all four source languages when translating into a language such as English, which does not normally attribute gender to inanimate objects. However, the pronoun error list should be extended to enable the detection of other pronominal errors, particularly involving relative pronouns (Table 6.8), in output from other source languages.

The algorithm developed for the detection of untranslated words and incorrect acronyms is the most successful method in terms of portability to other language pairs, detecting over 70% of errors in the three samples without making any changes to the English lexicon. However, in order to improve precision for any of the language pairs, steps must be taken during each evaluation to add named entities and new words to the lexicon as soon as they are identified.

Chapter 7

Conclusions and further research

Established metrics for the human evaluation of MT output, described in 3.4, are expensive and time-consuming, so interest has turned more recently to automated methods. As the majority of published automated methods, described in 3.5, still require human 'reference translations' for comparison, the primary goal of this research has been to investigate the potential of a previously unexplored approach that requires access only to the MT output.

An in depth pilot study has been carried out to investigate the potential of an MT evaluation method, which is based on the automated detection of errors in English MT output without the need for human translations or access to the source text. This investigation has involved a number of goals, which are summarised in 7.1 along with conclusions on the main achievements. In 7.1.1, achievements to date are compared with the list of desiderata for automated MT evaluation presented in 3.5.7.1. Findings from this research indicate several paths for future research; suggestions for further work are listed in 7.2. Final reflections are presented in 7.3.

7.1 Goals and achievements

1. *To design and compile a shareable corpus for MT evaluation to reflect the needs of MT users.*

Existing corpora for MT evaluation were either not publicly available or comprised texts or language pairs that did not suit our needs. The DARPA corpus (White et al., 1994) is publicly available but does not reflect the use of MT in the commercial world, as it includes only news texts. As a result, a new corpus was compiled specifically for MT evaluation based on a detailed design rationale (Elliott et al., 2003).

The rationale for compiling the corpus was based on two sources:

i. Responses to a world-wide survey of MT users, to identify the text types and language pairs translated most frequently by MT systems;

ii. An analysis of human scores from the DARPA corpus, to determine how many texts would be required to provide reliable evaluation results for competing systems.

In response to these findings a new, shareable parallel corpus was developed, specifically for French-English MT evaluation (Elliott et al., 2004a). French texts with existing English translations were selected from the web and copyright permission sought where necessary. All source texts and their human translations were meticulously checked for errors and translation correspondence, and corrections made where required. The corpus comprises the following:

- **40 original French texts**, each of approximately 400 words, in the computer software domain. Texts reflect commercial use of MT and comprise extracts from 10 user manuals, 10 Frequently Asked Questions, 10 press releases and 10 technical reports;
- **Four English machine translations** of each of the 40 French texts by competing systems;
- **One human translation** of each text for comparison in human evaluations of the texts;
- **One human 'reference translation'** of each text to enable the evaluation of fidelity.

The corpus, incorporating the human fluency and adequacy scores obtained for translations of twelve of the texts and MT error annotations, is freely available for research purposes by contacting the author. Data formats are described in 4.6.

2. *To obtain human judgements of MT output from the corpus, based on established evaluation methods, against which automated scores would be compared.*

The four machine translations and one human translation of twelve texts from the corpus were evaluated by human judges (Elliott et al., 2004a). Scores were obtained for fluency and fidelity at the segment level (normally a sentence or heading), following DARPA methodologies (White et al., 1994). Scores from three evaluators were obtained for each segment to reduce the effects of subjectivity. Judgements were obtained in electronic form

to facilitate the collation of scores for segments, texts, text types and systems. Human annotated scores for fluency and adequacy form an integral part of the corpus and are shareable for use in other MT evaluation projects.

Three significant conclusions were drawn from an analysis of human quality judgements:

i. Scores placed the MT systems in three different rank orders depending on which text type was being evaluated. This indicates that the evaluation of one text type cannot reliably indicate system performance for another text type and emphasises the need for texts that reflect a user's intended use of MT;
ii. A high correlation was observed between fluency and adequacy scores by text type and by system, indicating that an automated evaluation method, based on the analysis of the target text alone (like the human fluency evaluation conducted here), could well enable the prediction of human adequacy scores as well;
iii. Differences in system performance, according to human scores, were minuscule in some cases. This highlighted the anticipated difficulties in achieving automated scores that would reflect such close judgements.

3. *To develop a data-driven error classification scheme for the description of errors in French-English MT output with access only to the translation.*

An MT error classification scheme was required to enable the consistent manual annotation of errors throughout the MT output. This would ultimately guide the selection of errors for automated detection. All existing schemes were found to rely on access to the source text and many were based on errors found in translations out of English.

A data-driven approach was used to develop a fluency error classification scheme, during the author's manual annotation of errors in the 20,000 words of MT output, for which human scores had been obtained (Elliott et al., 2004b). No reference was made to the source text during this process, as automated error detection methods would be designed to reflect human thought processes with access to the MT output alone.

The resulting hierarchical fluency error classification scheme (Figure 4.16) enables the manual annotation of all errors found in French-English MT output, despite the lack of a source text or human translation for comparison. It also enables the annotation of any item

with more than one error type (inappropriate word, incorrect tense, incorrect position etc.) During the development of the scheme it became clear that adequacy (semantic) errors are also identifiable, despite the absence of the source text, due to their contribution to disfluency.

4. *To obtain information on error type frequencies from the manual annotation of errors in the MT output sample.*

The 20,000 words of MT output were manually annotated with error types (Elliott et al., 2004b) in accordance with the fluency error classification scheme. This would:

i. provide information on the frequency of each error type to guide the selection of errors for automated detection;
ii. facilitate the analysis of each error type in context, so that appropriate detection methods could be devised;
iii. ultimately enable the accuracy of each automated error detection algorithm to be tested by providing a measure for comparison;

A total of 4,697 errors were annotated in the sample, equal to 3.59 errors per segment (normally a sentence or heading). More errors were found in the user manuals than in any other text type. This was largely due to (a) the number of incorrectly ordered compound nouns which are frequent in these texts and (b) the large number of inappropriately translated verbs [eg. "safeguard", "record" or "register" (a file) instead of "save"]. The press releases contained the smallest number of errors (340 errors per text in comparison with 467 errors in the user manuals), again showing that evaluations based on one text type cannot represent system performance as a whole.

The most frequent errors were 'inappropriate' words, representing 42.2% of all annotated errors (see Table 4.9). Verbs were the most frequent part-of-speech in this category, comprising 25.4% of the total number of errors in this group. However, despite the large number of words labelled as 'inappropriate', in 88% of cases, the meaning was clear. The majority of these items were considered inappropriate due to incorrect word choices in the given context or to unnatural collocations, rather than incomprehensibility.

Incorrect word positions represented the second most frequent group of errors, comprising 19.2% of all annotated errors in the sample. The majority (64%) of these were incorrectly arranged compound nouns. Incorrectly positioned adverbs and adjectives were also relatively frequent, representing 12% and 11% respectively of errors of this type.

Prior to the development of automated detection methods, error frequencies were compared with human scores to explore whether any correlations existed. The total number of errors per system was compared with human fluency scores. A subset of errors considered to affect adequacy was then extracted and these frequencies were compared with human adequacy scores (see Table 4.10). Findings in both cases showed a rank correlation between error frequency and human scores by system. Furthermore, a very strong negative correlation was found between the two sets of values: for fluency $r = -0.998$ and for adequacy $r = -0.981$. Findings indicated that adequacy as well as fluency judgements could be captured based on the frequency of errors annotated without reference to the source text. However, no significant correlation was found between error frequencies and human scores by system at the level of text type (based on approximately 1,200 words).

5. *To develop rule-based methods, using human processing rationales, which detect particular error types in French-English MT output, without access to the source text or to any human translation.*

The sub-goals during the development of methods for automated error detection were as follows:

- Produce finely differentiated scores for output from competing systems;
- Detect as many errors as possible within each error group;
- Minimise the incorrect flagging of correct items;
- Where necessary, incorporate a feature to reflect human learning, which enables the user to enter new terms that should not be flagged as errors;
- Investigate the possibility of incorporating automated error correction.

Five error types were selected for automated detection. The aim in each case was to devise a method to reflect human processing concepts and rationales from initial supervised methods to develop a template for an intelligent system capable of fully-automated error

detection. Each algorithm was evaluated by comparing the number of automatically detected errors in each category with manual annotations.

i. Outrageous words

The MT output is compared against a lexicon of common mistranslations from French into English and all matches are flagged. The lexicon was generated by automatically comparing a 750,000 word sample of French-English MT output in the computing domain with a three million word corpus of comparable original English texts. English words found in the MT output but not in the original English texts were extracted and, after manual checks, an updateable 'outrageous' lexicon was created.

The method developed to detect 'outrageous' words, defined in 5.2.1, automatically identified all 52 errors of this type in the sample along with 13% of the remaining 1,367 content words that had been manually annotated as 'inappropriate'. No tokens were incorrectly identified as errors (Elliott et al., 2005).

ii. Inappropriate content words

The MT output is compared with an updateable list of unacceptable word combinations, each involving one inappropriate content word, found in a large sample of MT output in the same domain. Inappropriate compound nouns are not included, as their correct translations could be added at any stage to an MT system's user dictionary; focusing on other kinds of incorrect word combinations, which are not so easy to rectify, should make results from this method more robust in the long term.

Of the 1,189 remaining inappropriate content words that were not detected by the 'outrageous' word algorithm, 38% were detected using this method. No items were incorrectly flagged.

iii. Inappropriate prepositions

The MT output is compared against a list of inappropriate prepositions adjacent to their governing nouns, verbs or adjectives, found in a large sample of MT output in the same domain. Very infrequent word combinations are not included, nor are prepositional errors in word combinations involving other types of error. Only short entries were selected, most

often involving two or three words to maximise the number of matches found in the MT output.

Using the method developed for the automatic detection of inappropriate prepositions, 47% of the 323 manually annotated errors were automatically flagged. No items were incorrectly identified as errors.

iv. Inappropriate pronouns

The MT output was compared against a list of frequently occurring inappropriate anaphoric and relative pronouns. This was based on a detailed analysis of pronominal errors and, with knowledge of the French language, an understanding of why particular errors were regularly occurring. Due to the fact that French is a gender-based language, inanimate objects were often referred to as 'he' or 'she'. These words were marked as errors due to their proven rare usage in the selected text types, but this decision was the cause of the eight incorrectly flagged items.

Of the 104 manually annotated inappropriate pronouns, 69% were automatically detected using the developed method. Eight items were incorrectly flagged.

v. Untranslated words and incorrect acronyms

Untranslated words are detected by comparing the MT output against an updateable lexicon of tokens acceptable in English texts. The lexicon was compiled by creating a word list from the written part of the British National Corpus. Each time an evaluation is performed the lexicon should be updated with any acceptable new words, acronyms or named entities to reflect the human language learning process and to reduce the incorrect flagging of correct items.

Using the method developed for the automated detection of untranslated words and acronyms, 96% of the 203 manually annotated errors were detected (Elliott et al., 2005). In addition, a large number of words were incorrectly flagged, for the most part due to unrecognised named entities found in the MT output. However, due to the consistency of NEs found in translations of the same source texts by each system, this did not prevent the

correct ranking of systems in terms of the frequency of this error type. The method was also found to detect spelling errors.

A significant finding was that, in spite of low precision due to the large number of flagged named entities, results showed that our automated identification of untranslated words and acronyms reflects human error annotations far more closely than Systran's automatic flagging of words not found in system dictionaries. For the Systran output, our method achieved a precision of 44.1% and a recall of 100%. In comparison, Systran achieved only 14.6% for precision and 66.7% for recall based on the analyses discussed in 5.2.5.1.

Having investigated methods for the automated detection of five error types, major findings were as follows:

- A significant number of errors are automatically detectable in MT output using rule-based methods, when the source text or a human translation is not available. Of the 2,049 manually annotated errors for which detection methods were developed, 54% were automatically identified.
- A greater percentage of errors are automatically detectable when the algorithms do not rely on lexical context. All 'outrageous' words and 96% of untranslated words were detected, based on the identification of individual words.
- Fewer errors are detected when context is required, ie. when combinations of two or more words require matching. This is one reason why lower percentages of inappropriate content words, prepositions and pronouns are detected automatically.
- Erroneous context prevents the detection of many errors. Errors found in erroneous context are not listed in the rules files for automatic detection, due to reasons given in 5.2.2.1 and 5.2.3.1.
- For four of the error types, it is possible to enable automatic error correction or to offer an alternative suggestion after each error identified. This is possible when the method relies on lists of errors to identify. It is not possible to incorporate this feature in the detection method used for the majority of untranslated words, as most of these errors are identified when tokens are *not* found in a list.

6. *To identify correlations between automated scores based on the frequency of detected errors and human judgements.*

Error frequencies and human judgements of fluency and adequacy were compared by text type and by system to identify a subset of errors that, when automatically detected, would generate scores that correlate with human judgements.

In terms of individual error types, the strongest correlation was found between the number of automatically detected untranslated words in output from each system and human judgements of both fluency and adequacy for three of the four text types and for the whole sample.

However, the highest correlations with human scores for both fluency and adequacy were achievable when taking four of the detected error types into account: outrageous words, inappropriate content words, inappropriate pronouns and untranslated words/acronyms. Using the whole sample of 12 texts to compare the four systems, the correlation coefficient between automated error counts and human scores was -0.983 for fluency and -0.997 for adequacy. The detection of these four error types, along with several other combinations of errors, placed the four MT systems in rank order according to human adequacy judgements. Bearing in mind that the focus was on the detection of semantic rather than syntactic errors, this result was encouraging. It also proves the potential of an automated method that can predict human fidelity judgements without access to the source text or to a human translation.

Stronger correlations with human scores could be sought by devising methods for the detection of other error types: evidence has already shown in 5.3.2 that with the addition of algorithms for more error types, closer correlations with human scores can be achieved. However, it must be noted that the algorithms have so far been tested on a small data sample, due to the fact that all texts needed to be evaluated by humans and manually annotated with errors, which is very costly was time-consuming. Stronger correlations with human judgements or, at least, better predictions of the human ranking of MT systems across all of the individual text types, may be achievable with a larger sample.

7. *To determine the portability of methods to other language pairs in which the target language is English.*

The algorithms developed for the French-English language pair were tested on output from German, Italian and Spanish. The aim was to identify the extent to which methods would need to be adapted to detect a similar proportion of errors in output translated from other languages.

Findings showed that for the detection of outrageous words, inappropriate content words and inappropriate prepositions, the existing error lists compiled for French-English are not portable for use with MT from German, Italian or Spanish. A new list of errors would need to be compiled for each language pair, even though the target language is the same.

The detection of inappropriate pronouns was more successful, due to the fact that German, Italian and Spanish, in addition to French, are gender-based languages. As a result, some of the same pronominal errors occur in output from all four source languages when translating into English, which does not normally attribute gender to inanimate objects. However, the pronoun error list should be extended to enable the detection of additional pronominal errors that occur in output from other source languages.

The detection method for untranslated words and incorrect acronyms was identified as the most successful approach in terms of portability to other language pairs, detecting 72%, 79% and 90% of errors in output from Spanish, German and Italian respectively, without making any changes to the English lexicon.

7.1.1 Comparing achievements with desiderata for automated MT evaluation

After investigating existing automated evaluation methods and analysing their weaknesses in 3.5, a number of ideals for automated MT evaluation were brought to light. These were outlined in a list of desiderata in 3.5.7.1. This list is reproduced below, along with observations based on findings to date.

1. Automated scores should be predictive of human judgements at the text level and, preferably, also at the sentence level

No automated MT evaluation method is currently able to predict human judgements at the sentence level and all methods described in 3.5 require more than a single text. However, findings from our new approach (see 5.3.2) show that correlations between automated and human scores improve as more error types are automatically detected. This indicates that such a method has the potential to predict human scores at the text level, or with a small number of sentences, when more error types are detectable. Findings to date, based on only five error types, showed that the closest correlations were achieved when the whole sample was taken into account (approximately 5,000 words of MT output per system). However, strong correlations could also be seen for some of the individual text types, based on approximately 1,250 words (see Figures 5.12 and 5.13 in Chapter 5). Experiments with error weighting could also improve correlations with human judgements based on a smaller sample.

2. An ideal method should produce finely differentiated scores (rather than mere ranks) between competing translations, whether of very good or very poor quality

The current method is already able to produce finely differentiated scores for output from each MT system, based on the frequency of detected errors in only five different categories. This is in spite of a small number of errors not being detectable when the context is particularly bad, due to the fact that, as a rule-based approach to evaluation, some error types rely on correct context for automatic detection.

3. The method should produce reliable scores for MT and human translations

This research has focused on the detection of errors that typically occur in MT output, and it has been found that with the detection of only a small number of error types, strong correlations can be achieved with human judgements. However, these errors would rarely be made by humans, and certainly not by native speakers of the target language. For this reason, the current method is not suitable for the evaluation of human translations. That is not to say that an analysis of errors in human translations could not lead to the development of algorithms for automatic error detection.

4. It should be informative for users and developers

Unlike BLEU, which only produces a single score, the method developed here is informative for both users and developers. Strong correlations between automated scores (currently based on only five error types) and human judgements have shown that this approach could help users to compare the quality of output from different MT systems. Furthermore, the method is useful for post-editors who need to identify errors quickly, and some error types can be automatically corrected.

In addition, the method would enable developers to quickly identify particular error types and their frequencies, allowing them to prioritise areas for system improvement. Our approach was designed specifically for declarative and comparative black box evaluations, focusing on the quality of the output alone. However, this kind of automated error detection would still be useful for developers conducting glass box evaluations: regular evaluations of MT components are required prior to system release, before and after modifications are made. These evaluations are performed with access to the internal workings of the system and test suites are often used to evaluate the successes or failures of new translation rules. Our approach can be seen as an extension of a test suite, and can be used in a similar way: to enable developers to quickly compare the frequency of selected error types before and after internal changes are made.

5. The approach should be usable with different text types in different domains

The method has been explored with four different text types in a single domain. Our analyses of language use in different corpora have shown that some words (such as outrageous words) or word combinations are acceptable in some text types or domains and not in others. This indicates that while many of the error rules are completely portable, others are not. For this reason, different sub-sets of rules are needed for different text types and domains. This is an area for further development.

6. It should be easily replicable

The current approach is not easily replicable: the methods developed are based on extensive corpus analysis and required a great deal of time. However, this has provided a valuable insight for future directions and indicates a need to investigate the potential of automated methods for the compilation of error lists. This is discussed further in 7.2.

7. It should be adaptable for use with any language pair

Investigations in Chapter 6 showed that four of the error detection algorithms were not successful when tested on output from Italian, German and Spanish. This means that detailed analyses of output from other languages would be necessary to extend the rule files to include errors that typically occur in translations from other source languages. However, the algorithm for the detection of untranslated words was able to detect the majority of errors and would only require a small amount of adaptation to detect errors from other languages. Results and suggestions for optimising this method for other language pairs are described in 6.5.

8. Unlike methods such as BLEU, a new approach should aim to account for the level of quality of poorly translated sentence components (eg. by using weighting mechanisms) rather than assuming that all tokens in a text are of equal importance

Although error weighting has not been explored in the current research, the new approach has been designed in a modular way to enable such investigations and this will be a focus for future work.

9. Results should be comparable across evaluations

The method is objective, so scores from different evaluations using similar corpora would be comparable using the existing algorithms. However, the algorithms would need to be extended to include errors that occur in different text types and domains. Investigations of comparability across evaluations involving different text types is an area for future work.

7.2 Future work

The above findings and achievements indicate several paths for future research. These are listed under a series of themes below.

1. Investigation of more easily replicable methods

The methods developed here eliminate the need for human translations, which are expensive to produce, or access to the source text. However, the approach is still not resource-light: a great deal of time was spent, *inter alia*, analysing errors in the MT output,

compiling corpora for the development of error lists and preparing and testing the algorithms. This was to enable complete control during the supervised development of methods, to ensure that automatically detected errors would reflect the human annotations.

Work involved the compilation of a three million word corpus of user manuals and FAQs, translations by three MT systems of a 250,000 word corpus of French user manuals and FAQs, two corpora of press releases and technical reports, each of one million words, and an English lexicon generated from the BNC. A full list of corpora and word lists developed for this work is provided in Appendix 7. Manual text analysis was also extremely time-consuming. This involved scrutinising MT output to find error patterns, ensuring complete consistency in the annotation of errors in output from the competing systems, and the comparison of automatic and manual annotations.

The ultimate goal is to develop a freely available web-based system, incorporating algorithms for the detection of as many error types as possible, to achieve strong and consistent correlations with human judgements across different corpora. However, should anyone wish to replicate the methods described in this thesis, this would be extremely time-consuming.

In view of this, an important area for future work is an investigation of available tools and methods for faster corpus comparison. One such tool is Wmatrix (Rayson, 2003 and 2005) which enables automatic part-of-speech and semantic tagging as well as the creation of word frequency profiles for corpus comparison. This may be useful for the development of detection methods for other error types. In addition, an investigation of methods for the automation or semi-automation of error list creation is an important area for future work.

2. Investigation of errors in data-driven MT

The current error detection algorithms are based on the analysis and classification of errors in output from rule-based MT systems, due to the fact that no data-driven systems were available for our use. In fact, very few of these third generation MT systems are commercially available to date.

The causes of errors produced by data-driven MT systems are quite different from those found in output from rule-based systems, as discussed in 2.4.5. Consequently, although lexical and syntactic errors, omitted and unnecessary words are still found, the distribution and nature of these errors is likely to differ when compared with output from rule-based MT systems. Furthermore, while some problems, such as disambiguation, pronoun resolution and the mistranslation of idioms may be less frequent in data-driven MT output, there is a greater probability of grammatical errors.

Therefore, an important area for further work is to test the current algorithms on MT output from at least one data-driven system. The manual annotation of errors in this output would also be required to investigate the accuracy of the existing algorithms and, in particular, to identify errors which are not detected. This would guide the extension of rules for the detection of errors produced by data-driven MT systems and may call for the addition of further error types to our classification scheme

3. Extension of current error lists

In order to detect a greater percentage of errors (where possible) the existing error lists should be extended by extracting errors from larger corpora and, if possible, identifying methods for automating this process. In addition, the English lexicon should be extended to incorporate more named entities to reduce the number of tokens incorrectly flagged as 'untranslated'. This will involve the collation of existing lists of named entities for inclusion in the lexicon.

4. Detection of other error types

An important area for future work is the development of methods for the detection of additional error types, as:

a) this research has shown that stronger correlations with human scores are achievable when more error types are detected;
b) current methods are based on the quality of output from state-of-the-art systems, but systems will improve in the future and some of the detectable errors may cease to occur.

Methods for the detection of syntactic errors (incorrect word position, omitted and unnecessary words) should be the focus of the next stage of this work. This will involve the

use of machine learning techniques on annotated and part-of-speech tagged corpora. Attempts should also be made to identify methods that are more portable to other domains and language pairs.

5. Testing algorithms on other corpora

"MT evaluation research should be particularly wary of evaluation measures with parameters tuned to particular corpora. Such measures can overfit their objective function, and give misleading rankings on previously unseen corpora." (Turian et al., 2003) It is essential that we test the existing algorithms on other sets of MT output. Ideally, the first step would involve the use of the remaining 28 texts in the Fr-En computing corpus. However, due to the expense of human evaluations, human judgements are not available for these translations, so an investigation of correlations with human scores is not currently possible. However, the availability of these texts would enable us to compare the frequency of automatically detected errors in a comparable corpus. Subsequent work would involve the testing of algorithms on data from other domains where human judgements are available.

6. Error weighting or prediction of real error frequencies

In order to improve correlations with human scores, there are two possible areas for future work. Based on the number of errors detected in each category and comparisons with the number of manual annotations, predictions of actual errors based on this ratio may improve correlations with human judgements.

Another strategy would be to investigate the weighting of error types to enable close predictions of human scores. Due to the modular nature of the developed methods, it is possible to take the gravity of each error type into account by introducing weighted scores, unlike BLEU (Papineni et al., 2001) which assumes that all tokens in a text are of equal importance.

7. Development of a system for error detection in French-English MT output

The ultimate goal is to develop a freely available web-based system, enabling the evaluation of any sample of French-English MT output, and potentially other language pairs. This should allow the user to select error types for detection, view the detected errors

in context and have the option to view error corrections where this is possible. In addition, the system would generate error frequency statistics and a predicted human score.

8. Testing improved algorithms on output from other source languages

Research reported in Chapter 6 has shown that four of the five error detection methods developed, to date, are not immediately portable to other language pairs. This has highlighted the importance of developing methods that can detect errors in English from several source languages, where possible, and will influence approaches to the detection of other error types. It may be more efficient to use MT output from a number of languages to investigate error detection methods in the future.

7.3 Final reflections

The human facility to process and generate language is one of the most complex tasks we perform, and translation from one natural language into another is a difficult and error-prone implementation. It is, therefore, encouraging to conclude that results from this thesis demonstrate that simple algorithms, developed using a human processing template, can begin to automate such an intelligent process as machine translation error identification. Findings have provided valuable insights to guide subsequent phases of research in this previously uninvestigated area.

As a former student of translation studies, this work has enabled me to gain knowledge and experience in other areas including evaluation metrics, corpus and computational linguistics, computer programming and statistics. This has enabled me to work on other research projects during the course of my PhD and I have contributed to several publications on MT evaluation. A complete list of publications is provided in Appendix 8.

Bibliography

Akiba, Y., Imamura, K. and Sumita, E. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Akiba, Y., Sumita, E., Nakaiwa, H., Yamamoto, S. and Okuno, H.G. 2003. Experimental comparison of MT evaluation methods: RED vs. BLEU. In *Proceedings of MT Summit* 9, New Orleans, LA.

Albisser, D. 1993. Evaluation of MT systems at Union Bank of Switzerland. *Machine Translation* 8, pp 25-27. Kluwer: Netherlands.

Allen, J. and Hogan, C. 2000. Toward the development of a postediting module for raw machine translation output: a controlled language perspective. In *Proceedings of the 3rd International Controlled Language Applications Workshop*. CLAW 2000.

Allen, J. 2003. Post-editing. In Somers, H. (ed.) *Computers and translation: a translator's guide*, pp 211-244. Benjamins: Amsterdam.

Al-Sulaiti, L. and Atwell, E. 2004. Designing and developing a corpus of contemporary Arabic. In *Proceedings of the 6th TALC Conference*. Granada, Spain.

AltaVista Search Service. 1997. Eliminating the language barrier on the web. A new language translation service for web content now available on Digital's AltaVista Search Service. White Paper.

AltaVista Babelfish.
Available from: http://world.altavista.com/

Arnold, D. 1990. Text typology and machine translation: an overview. In *Proceedings of Translating and the Computer* 10. ASLIB: London.

Arnold, D., Sadler, L. and Humphreys, R.L. 1993. Evaluation: An Assessment. *Machine Translation* 8, pp 1-24. Kluwer: Netherlands.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., & Sadler, L. 1994. *Machine Translation: An Introductory Guide.*
Available from: http://clwww.essex.ac.uk/~doug/book/book.html

Arnold, D. 2003. Why translation is difficult for computers. In Somers, H. (ed.) *Computers and translation: a translator's guide*, pp 119-142. Benjamins: Amsterdam.

ATA (American Translators Association) Framework for Standard Error Marking. 2002. ATA Accreditation Program.
Available from: http://www.atanet.org/bin/view/fpl/12438.html

Atkins, S., Clear, J. and Ostler, N. 1992. Corpus design criteria. *Literary and Linguistic Computing* v7, pp 1-16.

Atwell, E., Demetriou, G., Hughes, J., Schiffrin, A., Souter, C. and Wilcock, S. 2000. A comparative evaluation of modern English corpus grammatical annotation schemes. *ICAME Journal* 24, pp 7-23.

Babych, B. 2004. Weighted n-gram model for evaluating machine translation output. In *Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics* (CLUK). University of Birmingham, UK.

Babych, B. and Hartley, A. 2004. Extending the BLEU MT evaluation method with frequency weightings. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Babych, B., Elliott, D. and Hartley, A. 2004. Calibrating resource-light automatic MT evaluation: a cheap approach to ranking MT systems by the usability of their output. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Babych, B., Elliott, D. and Hartley, A. 2004. Extending MT evaluation tools with translation complexity metrics. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING), Geneva, Switzerland.

Babych, B. and Hartley, A. 2004. Modelling legitimate translation variation for automatic evaluation of MT quality. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Babych, B., Hartley, A. and Elliott, D. 2005. Estimating the predictive power of n-gram MT evaluation metrics across languages and text types. In *Proceedings of MT Summit* 10, Phuket, Thailand.

BAF Corpus (Bi-text anglais-français)
Available from: http://rali.iro.umontreal.ca/

Baker, M., Francis, G. and Tognini-Bonelli, E. 1993. Corpus Linguistics and Translation Studies. Implications and Applications. *Text and Technology*: In honour of John Sinclair, Benjamins: Amsterdam.

Balkan, L. 1994. Test suites: some issues in their use and design. Test Suites for Natural Language Processing (TSNLP), University of Essex.
Available from: http://d-www.dfki.uni-sb.de/tsnlp/

Baroni, M. and Bernadini, S. 2004. BootCaT: Bootstrapping sorpora and terms from the web. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Bennett, S and Gerber, L. 2003. Inside commercial machine translation. In Somers (ed.) *Computers and Translation: A Tránslator's Guide*, pp176-190. Benjamins: Amsterdam.

Biber, D. 1993. Using register-diversified corpora for general language studies. *Computational Linguistics* 19, pp 219-241.

Bohan, N., Breidt, E. and Volk, M. 2000. Evaluating translation quality as input to product development. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC), Athens, Greece.

Brennan, S.E., Friedman, M.W. and Pollard, C.J. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. 155-162.

British National Corpus:
Available at: http://www.natcorp.ox.ac.uk/

Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L. and Roossin, P.S. 1990. A statistical approach to machine translation. *Computational Linguistics* 16, pp 79-85.

Callison-Burch, C. and Flournoy, R.S. 2001. A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Carroll, J.B. 1966. An experiment in evaluating the quality of translations. In Pierce, J. *Language and Machines: Computers in Translation and Linguistics*. Report by the Automatic Language Processing Advisory Committee (ALPAC). Publication 1416. National Academy of Sciences National Research Council.

Church, K.W and Hovy, E.H. 1993. Good applications for crummy machine translation. *Machine Translation* 8, 239-258. Kluwer: Netherlands.

Correa, N. 2003. A fine-grained evaluation framework for machine translation system development. In *Proceedings of MT Summit* 9, New Orleans, LA.

Corston-Oliver, S., Gamon, M and Brockett, C. 2001. A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics* (ACL). Toulouse.

Coughlin, D. 2003. Correlating automated and human assessments of machine translation quality. In *Proceedings of MT Summit* 9, New Orleans, LA.

Crook, M. and Bishop, H. 1965. Evaluation of Machine Translation. Final report, Institute for Psychological Research, Tufts University.

Culy, C. and Riehemann, S.Z. 2003. The limits of n-gram translation evaluation metrics. In *Proceedings of MT Summit* 9, New Orleans, LA.

Dabbadie, M., Hartley, A., King, M., Miller, K.J., El Hadi, W.M., Popescu-Belis, A., Reeder, R. and Vanni, M. 2002. A hands-on study of the reliability and coherence of evaluation metrics. In *Proceedings of the 5th ISLE Workshop on MT Evaluation, 3rd International Conference on Language Resources and Evaluation* (LREC), Las Palmas de Gran Canaria, Spain.

Darwin, M. 2001. Trial and error: an evaluation project on Japanese-English MT output quality. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Dauphin, E. and Lux, V. 1996. Corpus-based annotated test set for machine translation evaluation by an industrial user. In *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen, Denmark.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of HLT 2002, Human Language Technology Conference*, San Diego, California.

Dorr, B.J. 1994. Machine Translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4), pp 597-633.

Doyon, J., Taylor, K. and White, J. 1998. The DARPA Machine Translation Evaluation Methodology: Past and Present. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas* (AMTA). Philadelphia, PA.

Doyon, J., Taylor, K. and White, J. 1999. Task-based Evaluation for Machine Translation. In *Proceedings of MT Summit* 7. Singapore.

EAGLES MT Evaluation Working Group. 1996. EAGLES Evaluation of Natural Language Processing Systems. Final Report. EAGLES Document EAG-EWG-PR.2, ISBN 87-90708-00-8. Center for Sprogteknologi, Copenhagen.

European Association for Machine Translation (EAMT):
Available at: http://www.eamt.org/

ECI/MCI 1, European Corpus Initiative Multilingual Corpus
Available from: http://www.elsnet.org/resources/eciCorpus.html

Eckert, M and Strube, M. 2000. Dialogue acts, synchronizing units, and anaphora resolution. *Journal of Semantics* 17, pp 51-89.

Elliott, D., Hartley, A. and Atwell, E. 2003. Rationale for a multilingual corpus for machine translation evaluation. In Proceedings of CL2003: International Conference on Corpus Linguistics, Lancaster University, UK.

Elliott, D., Atwell, E. and Hartley, A. 2004a. Compiling and using a shareable parallel corpus for machine translation evaluation. In *Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora, 4th International Conference on Language Resources and Evaluation* (LREC), Lisbon, Portugal.

Elliott, D., Hartley, A. and Atwell, E. 2004b. A fluency error categorization scheme to guide automated machine translation evaluation. In R.E. Frederking and K.B. Taylor, (eds.) *Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA), Washington DC. Berlin-Heidelberg: Springer-Verlag, pp 64-73.

Elliott, D., Elliott, J., Atwell., E. and Hartley, A. 2005. Using corpora to automatically detect untranslated and "outrageous" words in machine translation output. In *Proceedings of CL2005: International Conference on Corpus Linguistics*, Birmingham University, UK.

English-French Aligned texts. Available from Knut Hofland, http://kh.hd.uib.no/tactweb/en-fr.htm

Esselink, B. 2000. *A Practical Guide to Localization*. Benjamins: Amsterdam.

Estival, D., Falkedal, K., Balkan, L., Dauphin, E., Meijer, S., Netter, K., Oepen, S. and Regnier-Prost, S. 1994. Analysis of existing test suites. Report to LRE 62-089 (D-WP1) Test Suites for Natural Language Processing (TSNLP), University of Essex.
Available from: http://d-www.dfki.uni-sb.de/tsnlp/

ET10-63 Parallel Corpus
Available from: http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html

FEMTI: A Framework for the Evaluation of Machine Translation in ISLE
Available at: http://www.isi.edu/natural-language/mteval/

Flanagan, M. 1994. Error Classification for MT Evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, MD.

FOLDOC: Free On-Line Dictionary of Computing.
Available from: http://foldoc.doc.ic.ac.uk/foldoc/index.html

Font Litjos, A and Carbonell, J. 2004. The Translation Correction Tool: English-Spanish user studies. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC), Lisbon, Portugal.

Forsbom, E. 2003. Training a super model look-alike: Featuring edit distance, n-gram co-occurrence, and one reference translation. In *Proceedings of the Workshop Towards Systematizing MT evaluation, MT Summit* 9, New , LA.

FreeTranslation by SDL
Available from http://www.freetranslation.com

Gamon, M., Aue, A and Smets, M. 2005. Sentence-level MT evaluation without reference translations: beyond language modeling. In *Practical Applications of Machine Translation, Proceedings of the 10th Conference of the European Association for Machine Translation,* Budapest, Hungary.

Google Translate
Available from: http://www.google.com/translate_t

Guerra Martinez, L. 2003. *Human translation versus machine translation and full post-editing of raw machine translation output.* MA Dissertation. Dublin City University.

Guessoum, A. and Zantout, R. 2001. A methodology for a semi-automatic evaluation of the lexicons of machine translation systems. *Machine Translation* 16, pp 127-149. Kluwer: Netherlands.

Guessoum, A. and Zantout, R. 2001. Semi-automatic evaluation of the grammatical coverage of machine translation systems. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Halliday, T.C. and Briss E.A. 1977. The evaluation and systems analysis of the Systran machine translation system. NTIS ADA 036.07.

Hansen, G. 1999. *Probing the process in translation – methods and results*. Copenhagen Studies in Language.

Heid, U. and Hildenbrand, E. 1991. Some practical experience with the use of test suites for the evaluation of Systran. In *Proceedings of the Evaluator's Forum*. Les Rasses. ISSCO, University of Geneva, Switzerland.

Helmreich, S. and Farwell, D. 1998. Translation Differences and Pragmatics-Based MT. *Machine Translation* 13, pp 17-39. Kluwer: Netherlands.

Hobbs, J.R. 1978. Resolving pronoun references. *Lingua* 44, pp 311-338.

House, J. 1977. *A model for translation quality assessment*. TBL Verlag Gunter Narr: Tübingen.

Hovy, E., King, M. and Popescu-Belis, A. 2002. Computer-aided specification of quality models for machine translation evaluation. In *Proceedings of the 5th ISLE Workshop on MT Evaluation, 3rd International Conference on Language Resources and Evaluation* (LREC), Las Palmas de Gran Canaria, Spain.

Hovy, E. 2003. Principles of context-based machine translation evaluation. *Machine Translation* 17, pp 43-75. Kluwer: Netherlands.

Hutchins, W.J. 1986. *Machine translation: past, present, future*. Ellis Horwood: Chichester. Halsted Press: New York.

Hutchins, W.J. and Somers, H.L. 1992. *An introduction to machine translation*. Academic Press: London.

Hutchins, W.J. 1995. Machine Translation: A Brief History. In E.F.K Koerner (ed.) *Concise history of the language sciences: from the Sumerians to the cognitivists*, pp 431-445. Pergamon Press: Oxford.

Hutchins, W.J. 2000. *Early years in machine translation*. Benjamins: Amsterdam.

Hutchins, W.J. 2002. Machine translation today and tomorrow. In *Computerlinguistik: was geht, was kommt*? Festschrift für Winfried Lenders, hrsg. Gerd Willée, Bernhard Schröder, Hans-Christian Schmitz, Sankt Augustin: Gardez! Verlag, pp159-162.

Hutchins, W.J. 2003. Commercial systems: the state of the art. In Somers, H. (ed.) *Computers and translation: a translator's guide*, pp 161-174. Benjamins: Amsterdam.

Hutchins, W.J. 2005. Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation*, v17, no 1-2.

Hutchins, W.J., Hartmann, W. and Ito, E. 2005 Compendium of Translation Software, Eleventh Edition, July 2005. Available from:
http://ourworld.compuserve.com/homepages/WJHutchins/Compendium-11.pdf

Ide, N. and Macleod, C. 2001. The American National Corpus: A standardized resource for American English. In *Proceedings of CL2001: International Conference on Corpus Linguistics*. University of Lancaster, UK.

INDEXA, L'annuaire Internet des Professionnels.
Available from: http://www.indexa.fr/

Isahara, H., Uchino, H., Ogino, S., Okunishi, T., Kinoshita, S, Shibata, S., Sugio, T., Takayama, Y., Doi, S., Nagano, T., Narita, M. and Nomura, H. 1994. Technical evaluation of MT systems from the developer's point of view: Exploiting test-sets for quality evaluation. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. Columbia, MD.

Isahara, H. 1995. JEIDA's test-sets for quality evaluation of MT systems – Technical evaluation from the developer's point of view. In *Proceedings of MT Summit 5*. Luxembourg.

ITR, International Translation Resources. 2006.
Available at: http://www.itr.co.uk

ITU, International Telecommunications Union. CRATER Parallel Spanish-French-English corpus. Available from:
http://www.comp.lancs.ac.uk/computing/research/ucrel/corpora.html

Jones, D.A. and Rusk, G.M. 2000. Toward a scoring function for quality-driven machine translation. In *Proceedings of 18th International Conference on Computational Linguistics* (COLING-2000). Saarbrucken.

Jordan, P.W., Dorr, B.J. and Benoit, J.W. 1993. A First-Pass Approach for Evaluating Machine Translation Systems. *Machine Translation* 8, pp 49-58. Kluwer: Netherlands.

Jurafsky, D. and Martin, J.H. 2000. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice-Hall: London.

Kaki, S., Yamada, S. and Sumita, E. 1999. Scoring multiple translations using character n-gram. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium* (NLPRS), Beijing, China.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Kilgarriff, A. BNC database and word frequency lists. Available from:
http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html#raw

Kilgarriff, A. 2001. Web as corpus. In *Proceedings of CL2001: International Conference on Corpus Linguistics*. Lancaster, UK.

King, M and Falkedal, K. 1990. Using test suites in evaluation of machine translation systems. In *Proceedings of the 18th International Conference on Computational Linguistics* (COLING). Helsinki, Finland.

King, M. 1996. On the notion of validity and the evaluation of MT systems. In H. Somers (ed.), *Terminology, LSP and Translation*, Studies in Language Engineering in honour of Juan C. Sager. Benjamins: Amsterdam.

King, M., Popescu-Belis, A. and Hovy, E. 2003. FEMTI: creating and using a framework for MT evaluation. In *Proceedings of MT Summit* 9, New Orleans, Louisiana.

Klein, J., Lehmann, S., Netter, K. and Wegst, T. 1998. DiET in the context of MT evaluation. In *Proceedings of KONVENS 98, 4th Conference on Natural Language Processing*, Bonn, Germany.

Knight, K and Chander I. 1994. Automated postediting of documents. In *Proceedings of the National Conference on Artificial Intelligence* (AAAI), Seattle.

Koehn, P. and Knight, K. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition, 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA.

Koehn, P., Och, F. J., and Marcu, D. 2003. Statistical phrase based translation. In *Proceedings of HLTNAACL.*

Koehn, P. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Koh, S., Maeng,J., Lee., J.L., Chae, Y.S. and Choi, K.S. 2001. A test suite for evaluation of English to Korean machine translation systems. In *Proceedings of Machine Translation Summit* 8, Santiago de Compostela, Spain.

Krauwer, S. 1993. Evaluation of MT Systems: A Programmatic View. *Machine Translation* 8, pp 59-66. Kluwer: Netherlands.

Krings, H.P. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Edited by G.S. Koby. Kent State University Press.

Kulesza, A. and Shieber, S.M. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, MD.

Language Weaver Press Release. 2005. *Language Weaver Increases Enterprise Scalability for its Statistical Machine Translation Software, Version 3.1 Release*. Available from: http://www.languageweaver.com/article_97.php

Lappin, S. and Leass, H. 1994. An algorithm for pronouminal anaphora resolution. *Computational Linguistics*, 20(4), pp 535-561.

Lavie, A., Sagae, K. and Jayaraman, S. 2004. The significance of recall in automatic metrics for MT evaluation. In R.E. Frederking and K.B. Taylor, (eds.) *Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA), Washington DC, USA. (Berlin-Heidelberg: Springer-Verlag). 134-143.

Lehmann, S., Oepen, S., Regnier-Prost, S., Netter, K., Lux, V., Klein, J., Falkedal, K., Fouvry, F., Estival, D., Dauphin, E., Compagnion, H., Baur., J., Balkan, L. and Arnold, D. 1996. TSNLP - Test Suites for Natural Language Processing. In *Proceedings of the 16th International Conference on Computational Linguistics* (COLING-96). Copenhagen.

Lehrberger, J. and Bourbeau, L. 1988. *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation*. John Benjamins: Amsterdam

Leusch, G., Ueffing, N. and Ney, H. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. In *Proceedings of MT Summit* 9, New Orleans, LA.

Levenshtein, V.I. 1965. Binary codes capable of correcting deletions, insertions and reversals. Doklady Akademii Nauk, USSR, Vol 163(4), pp 845-848.

Lin, C.Y. and Och, F.J. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004), Geneva, Switzerland.

Loffler-Laurian, A-M. 1983. Pour une typologie des erreurs dans la traduction automatique. *Multilingua* 2 (2), pp 65-78. Mouton Publishers.

Loffler-Laurian, A.M. 1996. *La Traduction Automatique*. Presses Universitaires du Septentrion, Lille.

Loper, E.D. and Bird, S. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop of Effective Tools and Methodologies for Teaching NLP and Computational Linguistics*. Philadelphia, PA.

Loper, E.D. and Bird, S. Natural Language Toolkit (NLTK). Available at: http://nltk.sourceforge.net/

Maier, E., Clarke, A. and Stadler, H.U. 2001. Evaluation of machine translation systems at CLS Corporate Language Services AG. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Mann, W.C. and Thompson, S.A. 1988. Rhetorical structure theory: toward a functional theory of text organization. *Text*, 8(3), pp 243-281.

Marrafa, P. and Ribiero, A. 2001. Quantitative evaluation of machine translation systems: sentence level. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit* 8, Santiago de Compostela, Spain.

McClure, S and Flanagan, M. 2003. Overcoming the MT Quality Impasse. Systran Language Translation Technologies: Report #29917. Available from : http://www.systransoft.co.uk/IDC/29917.htm

McEnery, A. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: University Press

McEnery, A. 1997. Multilingual corpora – current practice and future trends. In *Proceedings of Translating and the Computer* 19. ASLIB: London.

Miller, K. 2000. *The machine translation of prepositional phrases*. Unpublished PhD dissertation. Georgetown University, Washington, DC.

Miller, K.J. and Vanni, M. 2001. Scaling the ISLE Taxonomy: Development of metrics for the multi-dimensional characterisation of machine translation quality. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Minnis, S. 1993. Constructive Machine Translation Evaluation. *Machine Translation* 8, pp 67-75. Kluwer, Netherlands.

Minnis, S. 1994. A Simple and Practical Method for Evaluating Machine Translation Quality. *Machine Translation* 9, pp 133-149. Kluwer, Netherlands.

MULTEXT (Multilingual Text and Corpora). Available from: http://www.lpl.univ-aix.fr/projects/multext/

Nagao, M. 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. in A. Elithorn and R Banerji (eds) *Artificial and Human Intelligence*, pp 173-180. Amsterdam.

Nagao, M., Tsujii, J. and Nakamura, J. 1985. The Japanese Government project for machine translation. *Computational Linguistics* 11(2/3), pp 91-109.

Neal, J.G., Feit, E.L. and Montgomery C.A. 1993. Benchmark Investigation/Identification Project. *Machine Translation* 8, pp 77-84. Kluwer: Netherlands.

Nerbonne, J., Netter, K., Diagne, A.K., Klein, J. and Dickmann, L. 1993. A Diagnostic Tool for German Syntax. *Machine Translation* 8, pp 85-107. Kluwer: Netherlands.

Ng, V. and Cardie, C. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp 104-111. Philadelphia, PA.

Ng, V. 2004. Learning noun phrase anaphoricity to improve coreference resolution: Issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pp 152-159. Barcelona, Spain.

Niessen, S., Och, F.J., Leusch, G. and Ney, H. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC), Athens, Greece.

NIST 2005a. The 2005 NIST Machine Translation Evaluation Plan (MT-05). Available from: http://www.nist.gov/speech/tests/mt/doc/mt05-evalplan.v1.1.pdf

NIST 2005b. MT Evaluation Official Results. August 2005. Available from: http://www.nist.gov/speech/tests/mt/mt05eval_official_results_release_20050801_v3.html

Nomura, H. and Isahara, H. 1992. JEIDA's Criteria on Machine Translation Evaluation. In *Proceedings of the International Symposium on Natural Language Understanding and AI*, pp 107-117. Kyushu Institute of Technology, Iizuka, Japan.

Nyberg, E.H., Mitamura, T. and Carbonell, J.G. 1994. Evaluation metrics for knowledge-based machine translation. In *Proceedings of the 15th International Conference on Computational Linguistics*, pp 95-99, Kyoto, Japan.

O'Brien S. 2004. Machine Translatability and Post-Editing Effort : How do they relate? In *Proceedings of the 26th International Conference on Translating and the Computer.* ASLIB/IMI : London.

Och, F. J. 2002. Statistical Machine Translation: From Single-Word Models to Alignment Templates. PhD thesis, RWTH Aachen, Germany.

Och, F.J. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, (ACL), Sapporo, Japan.

O'Connell, T., O'Mara, F. and White, J. 1994. The ARPA MT evaluation methodologies: Evolution, lessons and further approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, MD.

O'Donnell, M. WagSoft Linguistic Software: Systemic Coder - a Text Markup Tool. Available from: http://www.wagsoft.com/Coder/index.html

Oepen S., Netter, K. and Klein, J. 1997. TSNLP - Test suites for natural language processing. In *Linguistic Databases*. CSLI Lecture Notes. CSLI: Stamford.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. Research Report RC22176. IBM Research Division. New York.

Papineni, K., Roukos, S. and Ward, T. 2002. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In *Proceedings of HLT 2002, Human Language Technology Conference*, San Diego, California.

Pastra, K. and Saggion, H. 2003. Colouring summaries BLEU. In *Proceedings of the Workshop on Evaluation Reuse, 10th Conference of the European Chapter of the Association for Computational Liguistics*. EACL03. Budapest, Hungary.

Petrits, A., Braun-Chen, F., Martínez García, J.M., Ross, C., Sauer, R., Torquati, A. and Reichling, A. 2001. The Commission's MT System: Today and Tomorrow. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Pierce, J. 1966. Language and machines: Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee (ALPAC). Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

Pinker, S. 1994. *The Language Instinct*. Penguin Books: London.

Popescu-Belis, A. 2003. An experiment in comparative evaluation: humans vs. computers. In *Proceedings of MT Summit* 9, New Orleans, LA.

Povlsen, C. and Bech, A. 2001. Ape: Reducing the monkey business in post-editing by automating the task intelligently. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Rajman, M. and Hartley, A. 2001. Automatically predicting MT systems rankings compatible with Fluency, Adequacy or Informativeness scores. In *Proceedings of the 4th ISLE Evaluation Workshop, MT Summit* 8, Santiago de Compostela, Spain.

Rajman, M. & Hartley, A. 2002. Automatic Ranking of MT Systems. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* LREC), Las Palmas de Gran Canaria, Spain.

Rayson, P. 2005. Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. Available from:
http://www.comp.lancs.ac.uk/ucrel/wmatrix/

Reeder, F. and Loehr, D. 1998. Finding the right words: an analysis of not-translated words in machine translation. In *Proceedings of 3rd Conference of the Association for Machine Translation in the Americas* (AMTA). Langhorne, PA.

Reeder, F., Miller, K., Doyon, K. & White, J. 2001. The Naming of Things and the Confusion of Tongues. In *Proceedings of the 4th ISLE Evaluation Workshop, MT Summit* 8, Santiago de Compostela, Spain.

Reeder, F. 2001. In one hundred words or less. In *Proceedings of the Fourth ISLE Evaluation Workshop, MT Summit* 8, Santiago de Compostela, Spain.

Reeder, F. and White, J. 2003. Granularity in MT evaluation. In *Proceedings of the Workshop Towards Systematising MT Evaluation, MT Summit* 9, New Orleans, Louisiana.

Reeder, F. 2004. Investigation of intelligibility judgments. In R.E. Frederking and K.B. Taylor, (eds.) *Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA), Washington DC, USA. (Berlin-Heidelberg: Springer-Verlag), 227-235

Resnik, P. 1998. Parallel strands: a preliminary investigation into mining the web for bilingual text. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas* (AMTA). Philadelphia, PA.

Resnik, P. 1999. Mining the web for bilingual text. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics* (ACL), University of Maryland.

Resnik, P. 2003. The web as a parallel corpus. *Computational Linguistics* 29 (3), pp 349-380.

Resnik, P. and Elkiss, A. 2005. The Linguist's Search Engine: Getting Started Guide. Technical Report LAMP-TR-108/CS-TR-4541/UMIACS-TR-2003-109, University of Maryland.

Richardson, S.D. 2004. Machine Translation of Online Product Support Articles Using a Data-Driven MT System. In R.E. Frederking and K.B. Taylor, (eds.) *Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA), Washington DC. Berlin-Heidelberg: Springer-Verlag.

Roudaud, B., Puerta, M.C. and Gamrat, O. 1993. A Procedure for the Evaluation and Improvement of an MT System by the End-User. In *Machine Translation* 8, pp 109-116. Kluwer: Netherlands.

SAE J2450: Translation Quality Metric. 2001. Society of Automotive Engineers, Warrendale, USA.

Sager, J.C. 1983. Quality and standards – the evaluation of translations. In C. Picken (ed.) *The Translator's Handbook.* Dorset Press

Shiwen, Y. 1993. Automatic evaluation of output quality for machine translation systems. *Machine Translation* 8, pp 117-126. Kluwer: Netherlands.

Simard, M. 1998. The BAF: A corpus of English-French bitext. In *Proceedings of the First International Conference on Language Resources and Evaluation*. Granada, Spain.

Simpkin, R.E. 1983. Translation Specifications. In C. Picken (ed.) *The Translator's Handbook.* Dorset Press

Sinaiko, H.W. and Klare, G.R. 1972. Further experiments in language translation: readability of computer translation. *ITL*, 15, pp 1-29

Sinaiko, H.W. and Klare, G.R. 1973. Further experiments in language translation: a second evaluation of the readability of computer translations. *ITL*, 19, pp 29-52.

Sinclair, J. 1991. *Corpus creation. Corpus, concordance, collocation*, pp 13-19. Oxford University Press.

Snell, B. and Crampton, P. 1983. Types of Translations. In C. Picken (ed.) *The Translator's Handbook*, pp 109-120. Dorset Press.

Somers, H. 1999. Review Article: Example-based Machine Translation. *Machine translation* 14, pp 113-157. Kluwer: Netherlands.

Somers, H. and Wild, E. 2000. Evaluating machine translation: the cloze procedure revisited. In *Proceedings of Translating and the Computer*. ASLIB: London.

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

Strube, M. 1998. Never look back: An alternative to centering. In *Proceedings of the 36th Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics*, v2, pp 1251-1257.

Strube, M. and Müller, C. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, (ACL), Sapporo, Japan.

Su, K.Y., Wu, M.W. and Chang, J.S. 1992. A new quantitative quality measure for machine translation systems. In *Proceedings of the 14th International Conference on Computational Linguistics* (COLING), Nantes, France.

Sugaya, F., Takezawa, T., Yokoo, A and Yamamoto, S. 1999. End-to-end evaluation in ATR-MATRIX: speech translation system between English and Japanese. In *Proceedings of Eurospeech* 99, pp 2431-2434.

Takezawa, T. 1999. Building a bilingual travel conversation database for speech translation research. In *Proceedings of the 2nd International Workshop on East-Asian Language Resources and Evaluation*, Taipei.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real

world. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (LREC), Las Palmas, Canary Islands, Spain.

Taylor, W. 1953. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly*, 30, pp 415-433.

Taylor, K. and White, J. 1998. Predicting what MT is good for: user judgments and task performance. In *Proceedings of 3rd Conference of the Association for Machine Translation in the Americas*. Langhorne, PA.

TELRI: Trans European Language Resources Infrastructure: TRACTOR archive. Available from : http://www.ids-mannheim.de/telri/cdrom.html

Tetreault, J. 1999. Analysis of syntax-based pronoun resolution methods. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp 602-605.

Thompson, H.S. 1991. Automatic evaluation of translation quality : outline of methodology and report on pilot experiment. In *Proceedings of the Evaluator's Forum*, pp 215-223. Switzerland : ISSCO.

Thompson, H.S. and Brew, C. 1996. Automatic evaluation of computer generated text: Final Report on the TextEval Project. Internal Report. Human Communications Research Centre, University of Edinburgh.

TIDES. 2005. Available at : http://projects.ldc.upenn.edu/TIDES/tidesmt.html

Tidhar, D. and Kussner, U. 2000. Learning to select a good translation. In *Proceedings of the 18th International Conference on Computational Linguistics* (COLING-2000). Saarbrucken.

Tillmann, C. 2003. A projection extension algorithm for statistical machine translation. In Collins, M. and Steedman, M., editors, *Proceedings of EMNLP*, pp 1–8.

Tomás, J., Mas, J.A. and Casacuberta, F. 2003. A quantitative method for machine translation evaluation. In *Proceedings of the Workshop on Evaluation Reuse, 10th Conference of the European Chapter of the Association for Computational Liguistics.* EACL03. Budapest, Hungary.

Turian, J.P., Shen, L. and Melamed, I.D. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit* 9, New Orleans, LA.

United Nations Parallel Text Corpus. Available from: http://morph.ldc.upenn.edu/Catalog/LDC94T4A.html

Vanni, M. 1998. Evaluating MT systems : Testing and researching the feasibility of a task-diagnostic approach. In *Proceedings of Translating and the Computer* 10. ASLIB: London.

Vanni, M. and Miller, K. 2001. Scaling the ISLE Framework: Validating tests of machine translation quality for multi-dimensional measurement. In *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit* 8, Santiago de Compostela, Spain.

Vanni, M. and Miller, K. 2002. Scaling the ISLE Framework: Use of existing corpus resources for validation of MT evaluation metrics across languages. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain.

Van Slype, G. 1979. Critical Methods for Evaluating the Quality of Machine Translation. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk.

Vasconcellos, M., Scott, B., Chandioux, J., and Tanaka, H. In Nirenburg, S (Ed). *1993 Forum: MT User Experience, in Progress in Machine Translation*. IOS Press Amsterdam. pp 243-252.

Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., and Waibel, A. 2003. The CMU statistical machine translation system. In *Proceedings of MT Summit IX.*

Voilà free online machine translation. Available at: http://tr.voila.fr/

Wagner, S. 1998. Small scale evaluation methods. In *Proceedings of the Workshop on the Evaluation of Machine Translation Systems, KONVENS* 98. Bonn, Germany.

Walker, M.A. 1989. Evaluating discourse processing algorithms. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, (ACL), University of Maryland.

Walker, D. 2004 Perspectives from Commercial Translation. Keynote speech, *6th Conference of the Association for Machine Translation in the Americas* (AMTA), Washington DC.

White, J., O'Connell, T. and O'Mara, F. 1994 The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the 1st Conference of the, Association for Machine Translation in the Americas*, Columbia, MD.

White, J. 1997. Tutorial slides on MT Evaluation. MT Summit 6. San Diego.

White, J. and Taylor, K. 1998. A task-oriented evaluation metric for machine translation. In *Proceedings of the 1st Conference on Language Resources and Evaluation* (LREC). Granada, Spain.

White, J., Doyon, J. and Talbott, S. 2000. Determining the tolerance of text-handling tasks for MT output. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (LREC), Athens, Greece.

White, J. 2000. Toward an automated, task-based MT evaluation strategy. In Maegaard, B., (ed.), Proceedings of the Workshop on Machine Translation Evaluation, *2nd International Conference on Language Resources and Evaluation* (LREC), Athens, Greece.

White, J. 2000. Contemplating Automatic MT Evaluation. In *Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, Cuernavaca, Mexico.*

White, J. and Forner, M. 2001. Predicting MT fidelity from noun-compound handling. In *Proceedings of the Fourth ISLE Evaluation Workshop, MT Summit* 8, Santiago de Compostela, Spain.

White, J. 2001. Predicting intelligibility from fidelity in MT evaluation. In *Proceedings of the 4th ISLE Evaluation Workshop, MT Summit* 8, Santiago de Compostela, Spain.

White, J. 2003. How to evaluate machine translation. In Somers, H. (ed.) *Computers and translation: a translator's guide*, pp 211-244. Benjamins: Amsterdam.

Wilks, Y. 1992. MT Contrasts between the US and Europe. In J. Carbonell et al. (eds) JTEC Panel Report commissioned by DARPA and Japanese Technology Evaluation Center, Loyola College, Baltimore, MD.

Wilks, Y. 1992. Systran: It obviously works, but how much can it be improved? In J. Newton (ed.) *Computers in Translation: A Practical Appraisal.* Routledge: London.

Worldlingo free online machine translation. Available at: http://www.worldlingo.com/

Yamada, K. and Knight, K. 2001. A syntax-based statistical translation model. In *Proceedings of the Conference of the Association for Computational Linguistics* (ACL), Toulouse, France.

Yang, J. and Lange, E. 2003. Going live on the internet. In Somers, H. (ed.) *Computers and translation: a translator's guide*, pp 191-210. Benjamins: Amsterdam.

Yasuda, K. et al. 2000. An automatic evaluation method of translation capability by DP matching using similar expressions queried from a parallel corpus. In IEICE Technical Report, NLC2000-63. Tokyo: Institute of Electronics, Information and Communication Engineers.

Yasuda, K., Sugaya, F., Takezawa, T., Yamamoto, S. and Yanagida, M. 2001. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Yokoyama, S., Kashioka, H., Kumano, A., Matsudaira, M., Shirokizawa, Y., Kodama, S., Ehara, T., Miyazawa, S. and Murata, Y. 2001. An automatic evaluation method for machine translation using two-way MT. In *Proceedings of MT Summit* 8, Santiago de Compostela, Spain.

Zens, R., Och, F. J., and Ney, H. 2002. Phrasebased statistical machine translation. In *Proceedings of the German Conference on Artificial Intelligence(KI 2002).*

Appendix 1

MT user questionnaire

Dear All

I hope that the following will be of interest to MT users throughout the world. As part of my PhD research in the School of Computing and the Centre for Translation Studies at the University of Leeds, England, I will shortly begin compiling a multilingual corpus, comprising source texts aligned with human and machine translations, to be used for the evaluation of MT systems. The text types, topics, source and target languages for inclusion in the corpus will be based on information gathered from this short questionnaire. Corpus content should, therefore, truly represent the kinds of texts that translation departments and companies translate most often with the aid of fully automatic MT systems.

All respondents will be granted free access to the corpus and to our MT evaluation results as soon as they become available.

The information you provide will be used solely for the purpose of research into corpus design.

Please disregard any questions you are unwilling or unable to answer. We thank you in advance for your time.

Name of Company:

Nature of Business:

Contact Name and Email:

Role within Company:

Q1/4:
Which fully automatic MT system(s) do you use? Please delete as appropriate.

Enterprise Translation Server (SDL)
LogoMedia Enterprise Solutions
Logo Vista X Multilingual/Pro
MLTS
Reverso Expert/Mac/Pro
Smart Translator
Systran Enterprise/Pro
Transcend Corporate (SDL)
Other (please specify):

Q2/4:
Which language pair(s) do you translate using MT?

Source language(s)	Target language(s)

Q3/4:
Please give details of the text types you translate with the aid of fully automatic MT systems:

Text type	Approximate monthly word count	Specific subject areas (if applicable)
Emails		
Memoranda		
Business letters		
Internal Company documents		
Tourist/travel information		
Web pages		
Newspaper articles		
Academic papers		
Calls for tender		
Patents		
Legal documents		
Scientific documents		
Technical documents		
Medical documents		
Instruction booklets		
User manuals		
Other (please specify)		

Q4/4:
For what purpose is the MT output used? Please delete as appropriate.
Gisting
Immediate circulation
Post-editing for internal use
Post-editing for publication or client

Thank you for taking the time to complete this survey.

Debbie Elliott
School of Computing
University of Leeds
Leeds LS2 9JT
Tel: 0113 3436818
Email: debe@comp.leeds.ac.uk

Appendix 2

Responses to MT user questionnaire

(January 2003)

NB. Responses in red were not analysed as they involve the use of CAT tools rather than MT

	Company name and contact details	Date rec'd	MT systems used	Language Pairs†	Texts and topics	Purpose*
1	Derek Lewis **Director** **Foreign Language Centre** **University of Exeter** D.R.Lewis@exeter.ac.uk	21/1/03	Systran Enterprise/Pro, Globalink Power Translator	En-De En-Fr En-Es En-It De-En Fr-En Es-En It-En	Web pages for education Newspaper articles for education Academic papers	1 3
2	Roisin Saunier **Freelance consultant** Language technologies and localisation sub__rosa@hotmail.com	21/1/03	Power Translator	De-En It-En Jp-En Pt-En	Emails Web pages Newspaper articles on polotics Academic papers on natural language processing, linguistics, human-computer interaction, anthropology	1
3	Hajdi Kosednar **Project Manager** **IOLAR** (Localisation of all IT texts, software, documentation, web pages etc.) hajdik@iolar.com	21/1/03	Currently evaluating systems	En-Sl En-Cr En-Se De-Sl De-Cr De-Se	Web pages 10% Instruction booklets 10% User manuals 10% Software localization 70%	3
4	Victor Dewsbery **Freelance translator** Berlin Germany translation@dewsbery.de	21/1/03	Linguatec Personal Translator 2000	De-En	Internal Company Documents Legal documents (mainly contracts) Technical documents	4
5	Henrik W. Gade **Documentation designer and translator** **Medtronic A/S** Copenhagen (Medical Equipment) Henrik.wilhelm.gade@medtronic.com	21/1/03	Systran Enterprise/Pro	En- De En-Fr En-Dn En-Sw En-Fi En-Gr En-Du En-It En-Es En-Pt	Internal Company documents: c. 3,000 wpm 5% User manuals (medical equipment in Urology, Gastroenterology and Neurology): 20,000 wpm 38% Software Strings: 30,000 wpm 57%	4
6	Ruth **Research Student** **Macquarie University** Australia ruth@ics.mq.edu.au	21/1/03	Systran Enterprise/Pro	Fr-En De-En	Emails Tourist/Travel information Web pages Newspaper articles Academic papers	1
7	John Smart **Managing Director** **SMART *Communications* Inc.** (Software Development) jsmart@smartny.com	21/1/03	Smart Translator	ContEn-CanFr Pt-CanFr ContEn-SA.Es Pt-SA.Es	Instruction booklets Technical user manuals and administrative manuals (Aerospace, Telecoms, Pharmaceutical, Oil and Gas, *Banking and Finance*, Insurance, Automotive, Power Plants, eGov) 1,000-1,000,000 pages per month	4

8	Xiuming Huang **MT Consultant** **SDL International** China Office Software Localization xhuang@sdlintl.com	21/1/03	SDL Enterprise Translation Server, SDL Transcend Corporate, Saidi E-way (Beijing)	En-Ch Ch-En (Simplified & Traditional)	Web pages 20% Newspaper articles 10% Calls for tender 10% Patents 10% Technical documents 20% Instruction booklets 5% User manuals 25%	4
9	John Davis Advisor to translation dept National Bank of Greece jdavis@nbg.gr	22/1/03	TM only	En-Gr Gr-En	Bank reports Internal documents Promotional materials	Info not provided
10	**Translation Service:** **MT Team** **European Commission** Alain Reichling (acting Head of Unit) Francine Braun-Chen (MT team leader) Cameron Ross (MT team member) Angelo Torquati (MT team member) alain.reichling@cec.eu.int francine.braun-chen@cec.eu.int cameron.ross@cec.eu.int angelo.torquati@cec.eu.int (sender)	22/1/03	EC Systran (purpose-built)	En-Du En-Fr En-De En-Gr En-It En-Pt En-Es Fr-Du Fr-En Fr-De Fr-It Fr-Pt Fr-Es De-En De-Fr Es-En Es-Fr Gr-Fr	Emails Internal documents (eg. policy documents, minutes, working papers related to EU activities) Calls for tender Legal documents (EU legislation) Scientific documents Technical documents Medical documents User manuals In 2002: 739,000 pages: 255,838 by the Translation Service (as a basis for a fully polished translation); 339,557 by other Commission directorates (administrators, largely for gisting); 90,694 by other EU Institutions and bodies; 53,122 by public administrations in the Member States. No MT of Web pages is currently provided.	1 2 3
11	Carlos Paz **Translator** **European Commission** Administration carlos.paz-carneiro@cec.eu.int	23/1/03	EC Systran	En-Es Fr-Es	Memoranda 20% Letters 5% Web pages 8% Newspaper articles 2% Calls for tender 5% Legal documents 60%	4
12	Romulo Henrique Gomes Marques **Freelance Translator** Sao Gabriel - RS – Brazil romulohmarques@bol.com.br	23/1/03	L&H Power Translator Pro	En-Pt	Medical books Nursing books 50,000 words per month	4
13	Eiichiro Takasaki **Freelance technical translator** Japan e-takasaki@mse.biglobe.ne.jp	23/1/03	PC Transer 2000	En-Jp Jp-En	Business letters 5% Scientific documents 15% Technical documents 80%	4
14	Juhani Reiman, **Managing Director** **Pasanet Oy** (Translation Company) Finland juhani.reiman@pasanet.fi	23/1/03	Linguatec Personal Translator, Systran Enterprise/Pro, Transmart	Fi-En En-Fr	Emails Technical documents User manuals	1 3
15	Monika Röthlisberger **Language Technologies Manager** **CLS Corporate Language Services AG** (Human translation, adaptation, MT services) monika.roethlisberger@cls.ch	23/1/03	Globaliser, DTS, Lexshop from Comprendium	De-En De-Fr En-De En-Fr Fr-En Fr-De	Emails – many 15% Memoranda – many 15% Business letters – many 15% Internal Company documents – many 15% Legal documents – some 4% Technical documents – some 4% Medical documents – few 1% Instruction booklets – very few 1% User manuals – many 15% many reports by financial analysts on the performance of markets/stocks etc. 15%	1 2 3

16	Julia Aymerich **Senior Computational Linguist in charge of MT** **Pan American Health Organization** aymericj@paho.org	24/1/03	Own proprietary software (PAHO MT System)	En-Es En-Pt Es-En Pt-En Pt-Es (soon) Es-Pt (soon)	Memoranda 2% Business letters 1% Internal company docs 12% Academic papers 2% Legal documents 1% Scientific documents 15% Technical documents 50% Medical documents 25% User manuals 2%	1 4
17	Giles Tilling **Translator** **EBNER** Industrieofenbau Heat treatment specialists Cocos (Keeling) Islands Australia tg@ebner.cc	24/1/03	Free online MT	Any language into En or De	Business letters 10% Web pages (1,000 per month) 90%	1
18	Warren McManus Freelance translator warren.mcmanus@fastwebnet.it	24/1/03	TM only (Déjà Vu)	It-En	Business letters 2,000 Internal Company documents 15,000 Newspaper articles 3,000 Academic papers 10,000 Legal documents 10,000 Finanical texts (annual and interim reports) 15,000	4
19	Antonio S. Valderrábanos **Director** **The Bits and Text Company** (Bitext) Linguistic Technology asv@bitext.com	24/1/03	Reverso Expert/Mac/ Pro	En-Es Es-En	Technical documents	Testing MT systems
20	Steve Richardson **Senior Researcher** **NLP group** **Microsoft Research** steveri@microsoft.com	24/1/03	MSR-MT: Microsoft Research's own MT system	En-Es En-De En-Fr En-Jp Ch(simp)-En Es-En Fr-En Fr-Es De-En Jp-En Ch(simp)-Jp	Technical documents Initially, 50 million words of online support articles, with perhaps another 1 million words per month additional User manuals Potentially tens of millions of words per year, although we haven't started doing this yet.	1 2 3 4
21	Lorraine Riach Translator Riach Translations Germany riach-translate@t-online.de	27/1/03	TM only (TRADOS)	De-En	Internal Company Documents 10% Legal documents 90%	2
22	Dr. Gerd Willée, Scientific Associate University of Bonn Dept. of CL willee@uni-bonn.de	27/1/03	Not stated	De-En En-De	Short examples for teaching purposes	Teaching
23	Anders Svensson, **Development Manager** **Explicon Svenska AB** Translation and Technical Documentation provider asn@explicon.se	29/1/03	Reverso Pro Systran Pro L&H Power Translator Self-developed translation software for Scandinavian languages	En-Fr En-De En-Es En-No En-Dn Sw-Fr Sw-De Sw-Es Sw-No Sw-Dn	Web pages 50% User manuals 50%	4
24	Tran Ngoc Tuan **PhD researcher** **Ho Chi Minh City University of Technology** Vietnam t-n-tuan@hcm.vnn.vn	30/1/03	Systran online	En-Vt Vt-En	Web pages 1000 words per month Academic papers 2000 words per month	3

25	Jennifer Brundage **NLP (MT) specialist** **SAP AG** Business software applications Walldorf Germany Jennifer.Brundage@sap.com	5/2/03	Logo Vista X Multilingual/ Pro Reverso Expert/Mac/ Pro Logos 7 Comprendium	De-En En-Es En-Fr En-Jp En-Ru En-Pt	Internal company documents (project-related docs, test cases, technical specifications) Technical documents User manuals Other: Customer notes (error handling, best practices, tips & tricks, etc) Training material (also used for training externals/ customers) Around 3 to 5 million words per year per language direction using MT. All business application areas are covered: Financial Accounting, Controlling, Logistics, Sales & Distribution, Human Resources, Customer Relationship Management...)	1 3 4

† Key to language pairs

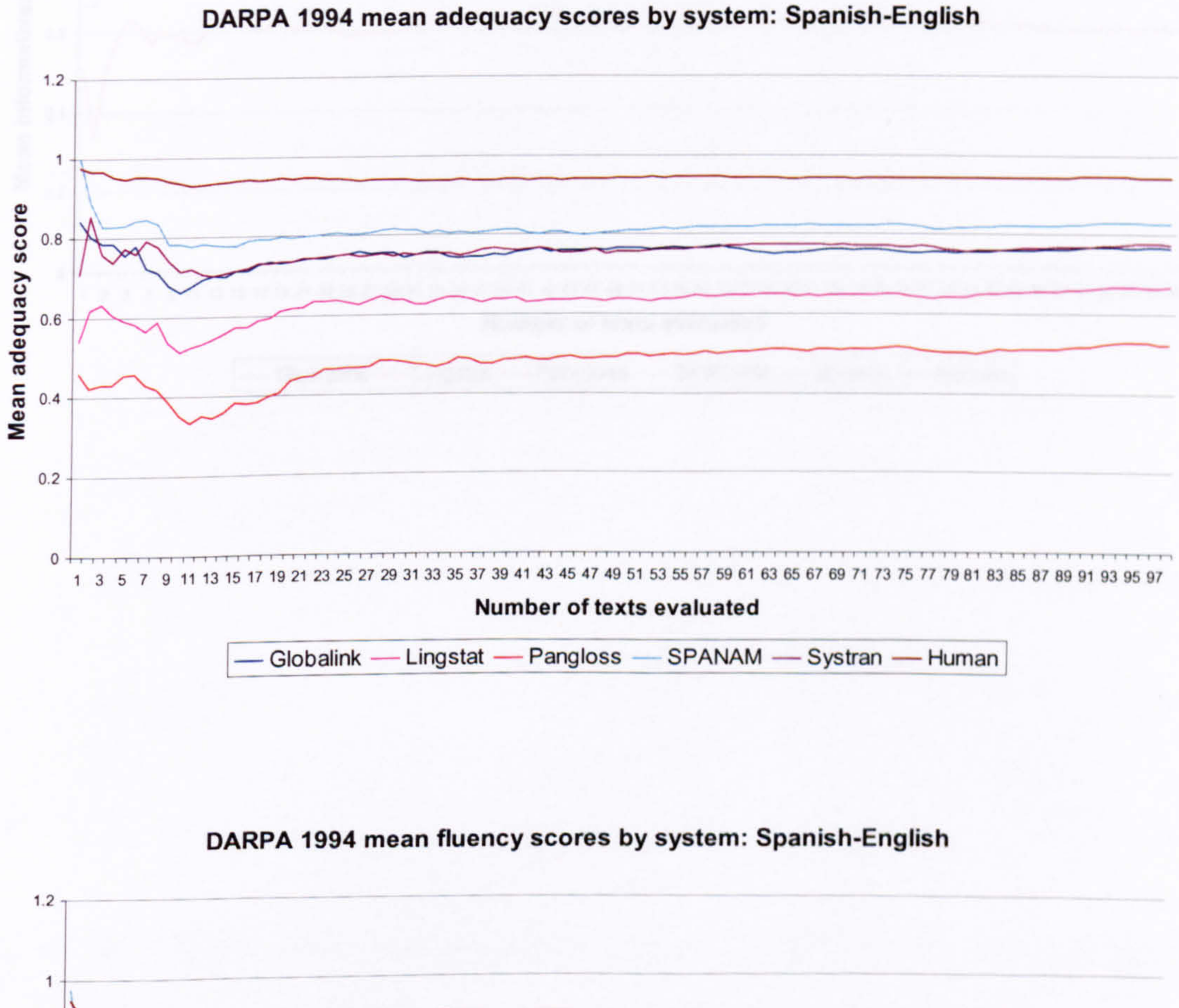
CanFr	Canadian French
Ch	Chinese
ContEn	Controlled English
Cr	Croatian
De	German
Dn	Danish
Du	Dutch
En	English
Es	Spanish
Fi	Finnish
Fr	French
Gr	Greek
It	Italian
Jp	Japanese
No	Norwegian
Pt	Portuguese
Ru	Russian
SA.Es	South American Spanish
Se	Serbian
Sl	Slovenian
Sw	Swedish
Vt	Vietnamese

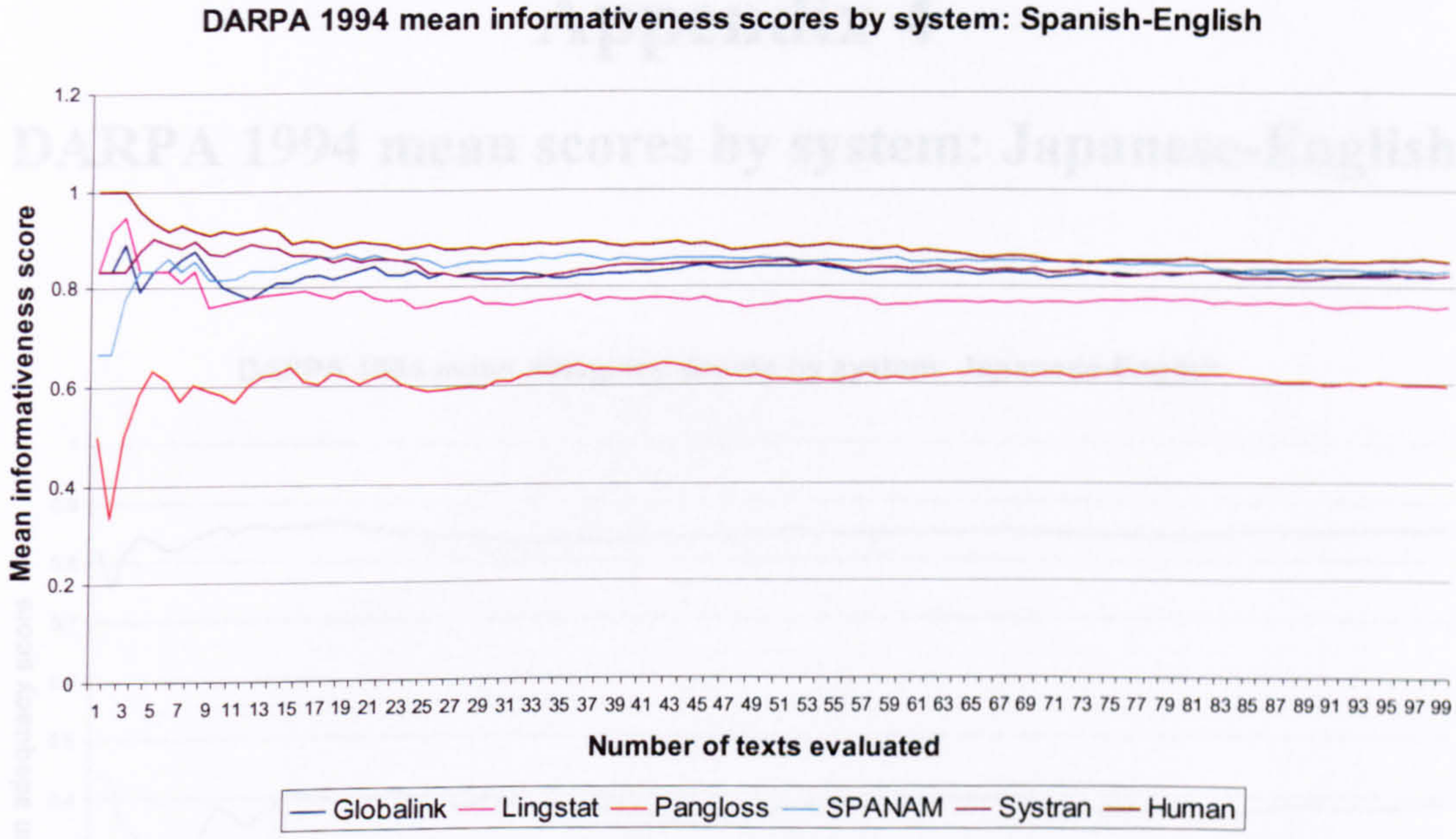
* Purpose of translation

1. Gisting
2. Immediate circulation
3. Post-editing for internal use
4. Post-editing for publication or client

Appendix 3

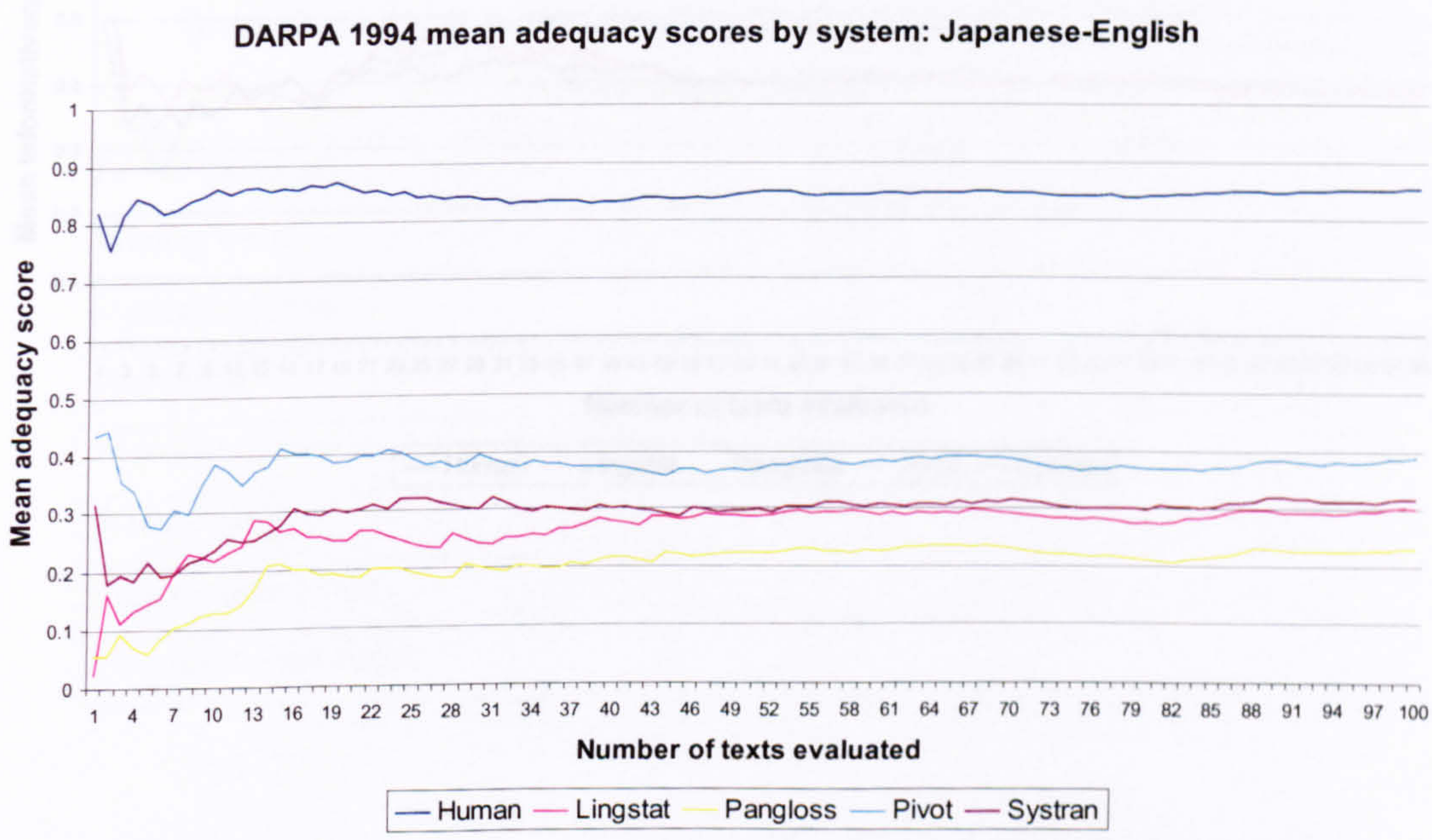
DARPA 1994 mean scores by system: Spanish-English

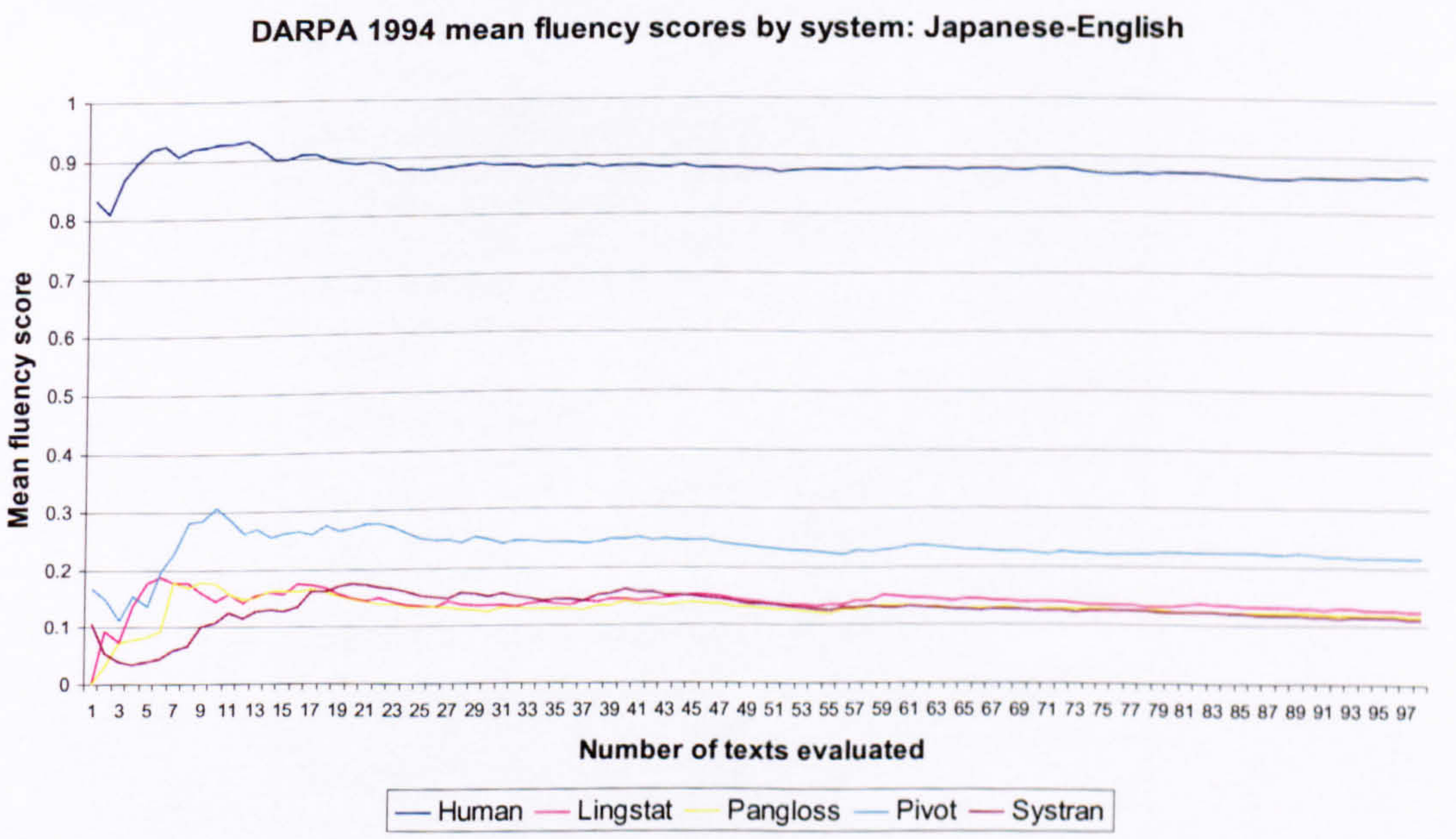


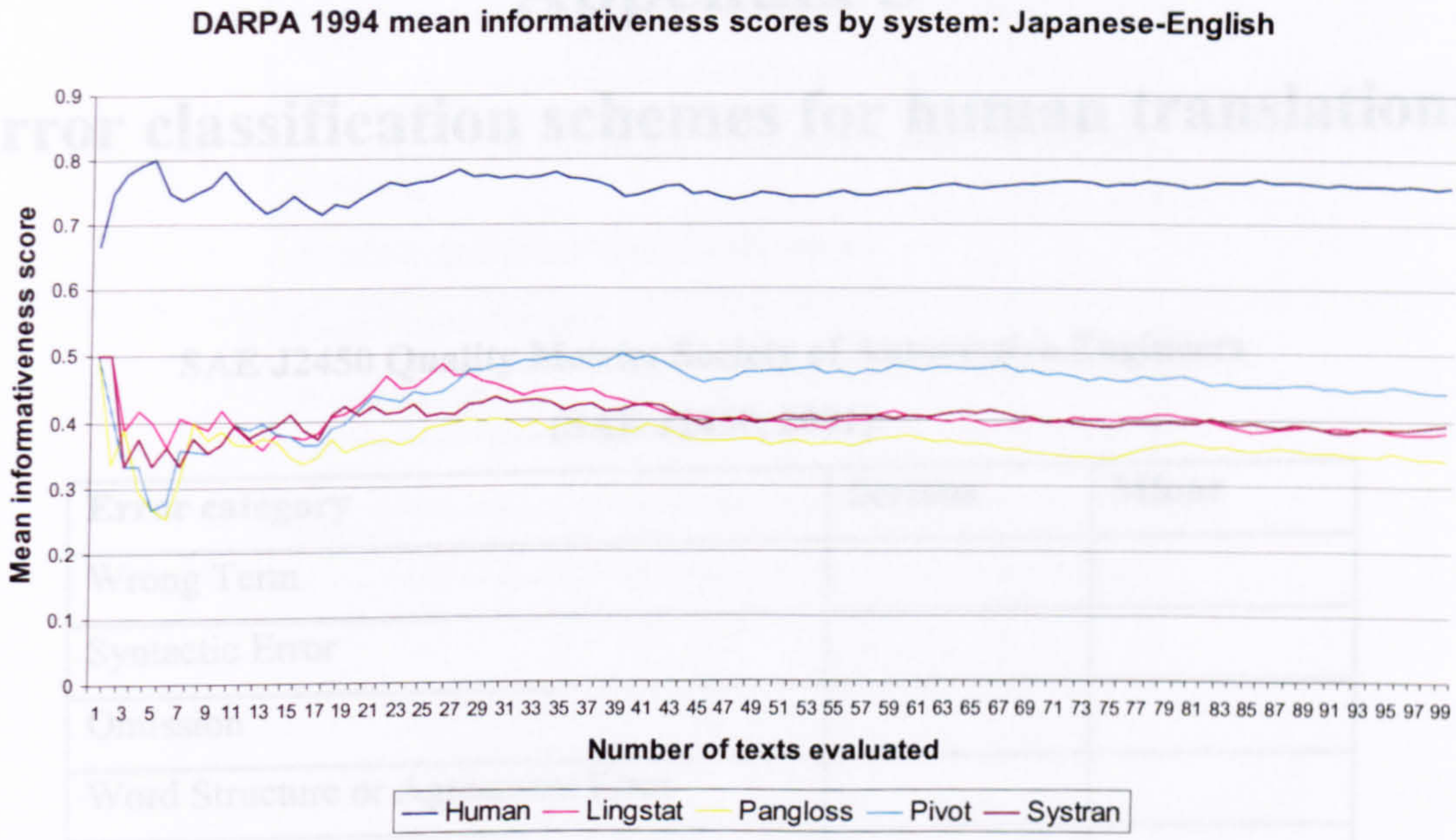

DARPA 1994 mean informativeness scores by system: Spanish-English
Mean informativeness score
1.2
1
0.8
0.6
0.4
0.2
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99
Number of texts evaluated
Globalink
Lingstat
Pangloss
SPANAM
Systran
Human

Appendix 4

DARPA 1994 mean scores by system: Japanese-English






DARPA 1994 mean informativeness scores by system: Japanese-English
Mean informativeness score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75 77 79 81 83 85 87 89 91 93 95 97 99
Number of texts evaluated
Human
Lingstat
Pangloss
Pivot
Systran

Appendix 5

Error classification schemes for human translations

SAE J2450 Quality Metric: Society of Automotive Engineers

(SAE J2450, 2001)

Error category	Serious	Minor
Wrong Term		
Syntactic Error		
Omission		
Word Structure or Agreement Error		
Misspelling		
Punctuation Error		
Miscellaneous Error		

Framework for Standard Error Marking: American Translators Association

(ATA, 2002) (Continued overleaf)

Incomplete passage
Illegible
Misunderstanding of original text
Addition or omission
Terminology, word choice
Register
Too freely translated
Too literal, word-for-word translation
False cognate
Indecision – gave more than one option
Inconsistency (same term translated differently)
Ambiguity

Grammar
Syntax
Punctuation
Spelling
Accents and other diacritical marks
Case (upper/lower)
Word Form
Usage
Style

BlackJack: ITR (ITR, 2006): extract from 21 error types

(The full list is commercially available as part of ITR's evaluation tool)

ERROR CODE	**Error type**
01	Misinterpretation of SLT
02	Non-application of glossary term
03	Wrong technical term in TLT
04	Inconsistent term in TLT
05	Discrepancy between TLT and SLT in embedded numerical value
06	Wrong treatment of acronym / of proper noun
07	SLT left untranslated

Appendix 6

Error classification schemes for machine translations

A typology of errors commonly found in automatic translation (Correa, 2003)

Feature	Description
IN	Input segment error
SEG	Segmentation error
TAG	Tag (markup) error
UW	Unknown word
NE	Named Entity error
AN	Source analysis error
TL	Target lexical error
TG	Target grammar error
TS	Target style error

Error categories for end-users (Wagner, 1998)

Wrongly translated word or expression

Incorrect order of words

Addition or removal of words

Untranslated word

Wrong punctuation

Typographical problems

Truncated word

Stylistically unhappy choice of words

Error Classification for MT Evaluation (Flanagan, 1994)

Category	Description
Spelling	Misspelled word
Not found word	Word not in dictionary
Accent	Incorrect accent
Capitalization	Incorrect upper or lower case
Elision	Illegal elision or elision not made
Verb inflection	Incorrectly formed verb, or wrong tense
Noun inflection	Incorrectly formed noun
Other inflection	Incorrectly formed adjective or adverb
Rearrangement	Sentence elements ordered incorrectly
Category	Category error (eg. noun vs verb)
Pronoun	Wrong, absent or unneeded pronoun
Article	Absent or unneeded article
Preposition	Incorrect, absent or unneeded preposition
Negative	Negative particles not properly placed or absent
Conjunction	Failure to reconstruct parallel constituents after conjunction, or failure to identify boundaries of conjoined units
Agreement	Incorrect agreement between subject-verb, noun-adjective, past participle agreement with preceding direct object, etc.
Clause boundary	Failure to identify clause boundary, or clause boundary unnecessarily added
Word selection	Word selection error (single word)
Expression	Incorrect translation of multi-word expression

Typology of errors to evaluate text type suitability for MT, based on linguistic problems for post-editors (Loffler-Laurian, 1996)

1	Vocabulary and terminology: Nouns Adjectives Verbs	**Errors based on meaning:** ▪ Incorrect translation of item (nonsense) ▪ Item appears correct but does not correspond to the original ▪ Unsuitable choice of item in lexical group (eg. in common collocation) ▪ Incorrect terminology in context **Errors based on usage** Item is comprehensible but does not conform to usage **Style** Choice of item not sufficiently elegant **Unknown vocabulary** Item left untranslated
2	Acronyms and proper nouns	Items remaining in source language or incorrectly translated
3	Prepositions	Incorrect prepositions in noun and verb phrases
4	Determiners	Incorrect translation, absence or unnecessary presence of articles, demonstratives and verb modifiers
5	Verb forms (tense)	Incorrect translation of tense (where tenses do not correspond in source and target languages)
6	Verbs (passive/active voice)	Unsuitable choice of passive or active voice, personal or impersonal expressions
7	Modal verbs	Unsuitable use of modal verbs
8	Negations	Translation of negative expression may be stylistically better if made positive in target language, and vice versa.
9	Word order	Word order may be grammatically incorrect or judged stylistically inferior
10	Arrangement	Long strings of successive elements (eg. descriptive proper nouns) are arranged incorrectly with added or unnecessary elements, or partly mistranslated so that the whole expression loses meaning

Error classification: Chaumier in (Van Slype, 1979)

- Noun phrase
- Subject and agent phrase (agent phrase = passive)
- Noun phrase and adjectival phrase
- Verbal phrase
- Verb phrases (object and adverbial phrases)
- Attribute (word or word group expressing the quality, nature, state of subject or object complement)
- Verb
- Negation
- Noun and noun phrase
- Articles
- Adjectives and adjectival phrases
- Pronoun and pronoun phrases
- Preposition and conjunction
- Constant words (invariable words such as proper nouns, abbreviations, figures)
- Punctuation

Error classification: Green in (Van Slype, 1979)

Structure errors: words and expressions in wrong order / incorrect attribution of adjectives / homograph errors (mainly from incorrect analysis of ST)
Preposition errors
Article errors
Errors in expressions: badly translated expressions
Translation errors: nouns, verbs and adjectives incorrectly translated
Miscellaneous errors: errors of number, misprints, superfluous words, foreign words treated as words of the SL

Appendix 7

Corpora and wordlists compiled for this research

Corpus	Description	Size
French-English MT evaluation corpus	French extracts from : 10 user manuals 10 FAQs 10 press releases 10 technical reports English MT output of all texts by 4 systems : Systran, Reverso, Comprendium, FreeTranslation English human translation of all texts English human 'reference translation' of all texts One human and four machine translations of 12 texts annotated with human fluency and adequacy scores by segment MT output of the same 12 texts annotated with error types	French source texts: 16,000 words English MT output: 64,000 words English human translations: 32,000 words
English corpus of software user manuals and FAQs	Ad hoc corpus comprising extracts mined from the web	3 million words of English
French-English corpus of machine translated software user manuals and FAQs	Ad hoc French corpus comprising extracts mined from the web English MT output of all texts by 3 systems : Systran, Reverso and Idiomax	French source texts : 250,000 words English MT output : 750,000 words
English corpus of computer software-related press releases	Ad hoc corpus comprising extracts mined from the web	1 million words of English
English corpus of computer software-related technical reports	Ad hoc corpus comprising extracts mined from the web	1 million words of English

German-English MT evaluation corpus	Extracts from : 3 user manuals 3 FAQs 3 press releases 3 technical reports English MT output from 2 systems : Systran and FreeTranslation	German source texts: 5,000 words English MT output: 10,000 words
Italian-English MT evaluation corpus	Extracts from : 3 user manuals 3 FAQs 3 press releases 3 technical reports English MT output from 2 systems : Systran and FreeTranslation	Italian source texts: 5,000 words English MT output: 10,000 words
Spanish-English MT evaluation corpus	Extracts from : 3 user manuals 3 FAQs 3 press releases 3 technical reports English MT output from 2 systems : Systran and FreeTranslation	Spanish source texts: 5,000 words English MT output: 10,000 words

Word list	Description	Size
Software lexicon	Word list extracted from the 3m word Corpus of English software user manuals and FAQs, enhanced with words from FOLDOC : the Free Online Dictionary of Computing	12,700 words
English lexicon	Word list generated from the written part of the BNC : British National Corpus	323,000 words
Outrageous word list	List of English words found in MT output of user manuals and FAQs from French but NOT in comparable original English texts	390 entries
Inappropriate content word list	List of word combinations involving inappropriate content words, selected from 90,000 words of MT output from French-English corpus of software user manuals and FAQs	1,490 entries
Inappropriate preposition list	List of inappropriate prepositions with their governing nouns, verbs, or adjectives, selected from 90,000 words of MT output from French-English corpus of software user manuals and FAQs	572 entries
Inappropriate pronoun list	Manually selected inappropriate pronouns found in French-English MT output	23 entries

Appendix 8

Author's publications

(in reverse chronological order)

1. Elliott, D., Elliott, J., Atwell., E. and Hartley, A. 2005. Using corpora to automatically detect untranslated and "outrageous" words in machine translation output. In *Proceedings of CL2005: International Conference on Corpus Linguistics*, Birmingham University, UK.

2. Babych, B., Hartley, A. and Elliott, D. 2005. Estimating the predictive power of n-gram MT evaluation metrics across languages and text types. In *Proceedings of MT Summit* 10, Phuket, Thailand.

3. Elliott, D., Atwell, E. and Hartley, A. 2004. Compiling and using a shareable parallel corpus for machine translation evaluation. In *Proceedings of the Workshop on The Amazing Utility of Parallel and Comparable Corpora, 4th International Conference on Language Resources and Evaluation* (LREC), Lisbon, Portugal.

4. Elliott, D., Hartley, A. and Atwell, E. 2004. A fluency error categorization scheme to guide automated machine translation evaluation. In R.E. Frederking and K.B. Taylor, (eds.) *Machine Translation: From Real Users to Research: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas* (AMTA), Washington DC. Berlin-Heidelberg: Springer-Verlag, pp 64-73.

5. Babych, B., Elliott, D. and Hartley, A. 2004. Calibrating resource-light automatic MT evaluation: a cheap approach to ranking MT systems by the usability of their output. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

6. Babych, B., Elliott, D. and Hartley, A. 2004. Extending MT evaluation tools with translation complexity metrics. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING), Geneva, Switzerland.

7. Hartley, A., Babych, B. and Elliott, D. 2004. Using corpora to evaluate machine translation. In Kimble, I (ed.) *Using corpora and databases*, pp 38-57. University of Portsmouth.

8. Elliott, D., Hartley, A. and Atwell, E. 2003. Rationale for a multilingual corpus for machine translation evaluation. In Proceedings of CL2003: International Conference on Corpus Linguistics, Lancaster University, UK.

9. Elliott, J. and Elliott, D. 2003. The Human Language Chorus Corpus (HULCC). In Proceedings of CL2003: International Conference on Corpus Linguistics, Lancaster University, UK.