An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness, and Acceptability of Machine-Translated Technical Documentation for French and German Users

A dissertation submitted to Dublin City University in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johann Roturier M.A.

School of Applied Languages and Intercultural Studies

November 2006

2 Volumes (Volume 1)

Supervisors: Prof. Jenny Williams, Dr Minako O'Hagan, & Dr Sharon O'Brien

Declaration

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

Signed: _____   Student Number: 52147321

(Candidate)

Date: _____29-01-07_____

Acknowledgements

There are many people to whom I am indebted due to their contribution to this dissertation over the last three years. First of all, I would like to thank my academic supervisors, Prof. Jenny Williams, Dr. Minako O'Hagan, and Dr. Sharon O'Brien, not only for their invaluable advice and recommendations, but also for their encouragement and eternal patience. I am extremely grateful to them for pushing me to explore new horizons and tackle topics from a different angle. Thanks to them, the experience has been all the more diverse and enjoyable.

I also wish to thank my industrial supervisor, Dr. Fred Hollowood, for his availability, advice, and sense of direction. On numerous occasions, he has helped me address what seemed like unsolvable 'level 9' problems. I am extremely grateful to him for providing me with resources when I needed them. Without him, his team in Dublin, and Symantec in general, this research could not have been carried out.

I am also grateful for the support I have received from all Symantec staff who have helped during these three years. Without Symantec's financial support, without their authorisation to allow me to access some of their resources and users, this dissertation would never have come to life. To all Symantec employees who have involved in one way or another in this research, I say a huge "Thank you!"

I would like to acknowledge the advice received from Dr. Sabine Lehmann from Acrolinx at the early stages of this research. Thank you very much for stimulating discussions on test suites and everything else!

I would also like to thank my family and friends for always supporting me despite the distance, and putting up with my prolonged silence when I was writing up this dissertation.

Last but not least, to Dympna, without whom I simply would not have completed this dissertation. Regardless of the time when things were good or bad, you have always been there to help me and support me, and this will never be forgotten.

Abstract

Johann Roturier: An investigation into the impact of Controlled English rules on the comprehensibility, usefulness, and acceptability of machine-translated technical documentation for French and German users.

Previous studies suggest that the application of Controlled Language (CL) rules can significantly improve the readability, consistency, and machine-translatability of source text. One of the justifications for the application of CL rules is that they can have a similar impact on several target languages by reducing the post-editing effort required to bring Machine Translation (MT) output to acceptable quality. In certain situations, however, post-editing services may not always be a viable solution. Web-based information is often expected to be made available in real-time to ensure that its access is not restricted to certain users based on their locale. Uncertainties remain with regard to the actual usefulness of MT output for such users, as no empirical study has examined the impact of CL rules on the usefulness, comprehensibility, and acceptability of MT technical documents from a Web user's perspective. In this study, a two-phase approach is used to determine whether Controlled English rules can have a significant impact on these three variables. First, individual CL rules are evaluated within an experimental environment, which is loosely based on a test suite. Two documents are then published and subject to a randomised evaluation within the framework of an online experiment using a customer satisfaction questionnaire. The findings indicate that a limited number of CL rules have a similar impact on the comprehensibility of French and German output at the segment level. The results of the online experiment show that the application of certain CL rules has the potential to significantly improve the comprehensibility of German MT technical documentation. Our findings also show that the introduction of CL rules did not lead to any significant improvement of the comprehensibility, usefulness, and acceptability of French MT technical documentation.

List of Figures

List of Tables

Abbreviations

AECMA	Association Européenne des Constructeurs de Matériel Aérospatial
AI	Artificial Intelligence
CASL	Controlled Automotive Service Language
CL	Controlled Language
CLAW	Controlled Language Application Workshop
CMS	Content Management System
CSS	Cascading Style Sheet
CTE	Caterpillar Technical English
GIFAS	Groupement des Industries Françaises Aéronautiques et Spatiales
GMS	Globalisation Management System
GUI	Graphical User Interface
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IDE	Integrated Development Environment
LGP	Language for General Purpose
LISA	Localisation Industry Standards Association
LSP	Language for Specific Purpose
MT	Machine Translation
NLP	Natural Language Processing
NP	Noun Phrase
PDF	Portable Document Format
PE	Post-Editing
POS	Part Of Speech
PP	Prepositional Phrase
RBMT	Rule-Based Machine Translation
RSS	Rich Site Summary
SE	Simplified English
TM	Translation Memory
UI	User Interface
URL	Uniform Resource Locator
VP	Verb Phrase
W3C	World Wide Web Consortium
XML	eXtensible Markup Language
XSL-FO	eXtensible Style Language Formatting Objects
XSL-T	eXtensible Style Language Transform

Introduction

Introduction

In 2003 this researcher was involved in a Machine Translation (MT) evaluation project from which the idea for this dissertation originated. This project was conducted within the localisation department of a software publisher, Symantec, which specialises in security and availability solutions. During this project, special attention was to be paid to the way source content was authored so as to reduce the Post-Editing (PE) of MT output to a minimum (Roturier, 2004). Such an objective was based on a long series of initiatives that have been undertaken in the field of technical communication, whereby publishers have attempted to increase the comprehensibility of their source content for both humans and machines by implementing a Controlled Language (CL). Huijsen (1998: 2) gives the following definition of a CL:

> A CL is an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style.

This definition can be supplemented by the fact that it is almost always used to 'write clear technical documentation in a particular domain'. (Power et al., 2003: 115). Industry-based projects (such as Bernth: 1998, Kamprath et al.: 1998, Godden: 1998, or Rychtyckyj: 2002) have shown that the quality of MT output can be significantly improved by removing complexity and ambiguity from source technical documentation. By placing strict linguistic and pragmatic restrictions on the source text, the quality of MT output can be improved to such an extent that it seems possible to significantly reduce translation turnaround and costs. Using this approach, the traditional translation process may be effectively replaced by a PE step to bring the quality of MT output to an acceptable standard. This has been empirically confirmed by O'Brien (2006: 177), who discovered that 'controlling the input to MT leads to faster post-editing'.

Until now, however, no research has been conducted to determine whether the application of CL rules can have a positive impact on the reception of machine-translated technical documents that are not post-edited prior to their publishing and dissemination to users. The aim of the present dissertation is to address this question and make a useful contribution to the field by focusing on the reception of machine-translated technical documents from a Web user's perspective.

Objectives of the study

The present study has two main objectives. First, it will attempt to determine whether some individual CL rules are more effective than others in improving the comprehensibility of MT output in two target languages when no post-editing is performed. If this were the case, the most effective rules could be used as authoring standards when developing specific Web-based technical documentation. In this study, the 'comprehensibility' of a translation is based on the definition provided by T.C. Halliday, who states that comprehensibility refers to 'the ease with which a translation can be understood, its clarity to the reader.' (quoted in Van Slype, 1979: 62). In order to achieve this objective, the following questions will have to be answered: are specific rules always effective regardless of the target language? Are there any cases where the rule may have a negative impact on the comprehensibility of the target text? In these cases, should the scope of the rules be reduced to allow for exceptions?

The second objective of this study is to investigate whether the application of certain CL rules can influence the reception of machine-translated documents by a specific category of information consumers: users of software products who require technical information in their own language. The decision to solicit feedback from real users stems from the general lack of involvement of users in the evaluation process of translations as end products (Lauscher, 2001: 166). It appears important to fill this gap since 'user feedback can make a valuable contribution to an assessment programme; the aim of which is to evaluate the quality of documentation after it is distributed' (Marlow, 2005: 36). The ease in accessing users of Symantec products is therefore a unique opportunity to address this question by involving users in the evaluation process of machine-translated documents within the framework of a case study.

In order to achieve this second objective, the following question will have to be answered. From a Web user's perspective, can the application of CL rules significantly improve the comprehensibility, usefulness, and acceptability of online machine-translated documents? In this study the term 'usefulness' is based on the definition provided by Richardson (2004: 247), whereby 'customers feel that an

article helped them solve their problem'. Several studies conducted at Microsoft and Cisco (Richardson, 2004, Jaeger 2004) have indeed suggested that refined MT output could be useful to users. The term 'refined MT output' is used to reflect any customisation an MT system may have received to be used in a specific domain without post-editing. Such customisation may include the use of specific user dictionaries or the design of extra rules. As mentioned by Allen (2003: 298), however, it is also 'important to determine to what extent MT output texts are acceptable' for their users. In this study, the term 'acceptability' is based on the fourth standard of textuality defined by De Beaugrande and Dressler (1981: 7). This standard concerns 'the text receiver's attitude that the set of occurrences should constitute a cohesive and coherent text having some use or relevance for the receiver'. Since this definition of acceptability contains the term 'use', the concepts of acceptability and usefulness must be further clarified. Acceptability does not only refer to the relevance a text has for its receiver, but also to the manner in which its textual characteristics are going to be accepted, tolerated, or rejected by its receivers. Lassen (2003: 76) infers that if a document does not meet specific expectations in terms of style, layout, and content, readers who belong to the discourse community 'would consider it unacceptable as a specimen of the genre'. In the context of software technical support documentation, the discourse community encompasses users of software products who need to find solutions to their problems. The machine-translated solutions some users are being provided with may be useful in certain cases, but are these documents acceptable as specimens of the genre? One possible way to answer this question is to determine whether users are willing to consult such documents again in the future. Based on De Beaugrande and Dressler's terminology (1981: 8), this will allow us to examine whether the potential 'disturbances' contained in machine-translated documents are being 'tolerated' by users of technical support documentation.

In this study the choice of the target languages was based on the ability of this researcher to analyse MT output, thus French and German were selected. It was important to select two languages that were not too similar, such as two Romance languages. It was also decided to focus on the output of a single MT system rather than to compare the impact of the rules on the outputs produced by several MT systems. This choice was prompted for several reasons. It was felt that the effort

involved in adding corpus-specific terminology to several systems' user dictionaries would be too great. Since commercial MT systems rely on proprietary terminology formats, the interchange of user dictionary resources from one MT system to the next is not automatically possible. Besides, evaluation costs could spiral if more than one MT system were used, since the time required for the evaluation process would increase significantly. Instead of having to reduce the amount of evaluation material, it was preferable to use only one MT system. Systran Web Server 5.0 was selected for a number of reasons. First, this system is used to power most Web-based translation portals today. Second, it is a broad coverage MT system so all sentences will be parsed in one way or another. Finally, this product was available in Symantec's localisation department. Symantec's user dictionaries of this MT system contain domain-specific terminology, thus reducing the terminological customisation required prior to the evaluation process.

Context and motivations

The benefits of CLs abound in the marketing documentation of certain CL checker providers (Smart, 2006: 8). A CL checker may be defined as an application designed to flag linguistic structures that do not comply with a predefined list of formalised CL rules. However, one should not forget that CLs may also have side-effects, since the acceptability of the source text could be affected by the introduction of rules. This danger is expressed by Huijsen (1998: 12), who states that 'some writing rules may even do more harm than good'. This problem was foreseen during the implementation of Caterpillar Technical English (CTE) at Caterpillar, but internal study findings revealed that CTE versions of source documents were well received by English readers. Not only were Caterpillar's translation costs reduced with the introduction of CTE, but the acceptability of source texts was also enhanced (Hayes et al., 1996: 87). Focus group testing revealed that the CTE versions of sentences were easier to understand, while the CTE versions of documents were easier to skim thanks to the low incidence of dependent sentences. This example confirms that a careful selection of specific CL rules is essential in achieving well-defined objectives.

However, deploying a large set of CL rules is sometimes not feasible due to time constraints. Having to check that a text conforms to a large set of CL rules will take time to complete, even with the use of a CL checker. In the worst case scenario, such a pre-editing process might take longer that the actual post-editing of the MT output. For instance, Govyaerts (1996: 139) reports that the controlled authoring process at Alcatel Bell could take up to 20% longer than the traditional process, thus affecting the writers and the whole document production workflow. Identifying such pitfalls was one of the objectives mentioned by Schachtl (1996: 143) during the design of a German CL for Siemens documentation. Siemens saw the benefits of using CL to improve MT, but they did not want to 'simplify the source text too much, obstruct the process of authoring and cause the texts to get too long' (ibid). Based on these recommendations, the impact of individual rules should therefore be evaluated empirically to determine whether the impact of a given rule justifies its implementation in a systematic way. However, Nyberg et al. (2003: 276) remark that 'the motivation for individual writing rules is generally based on intuition rather than on empirical evidence'. This may be due to the fact that determining the role that a specific individual CL rule plays in a given rule set is a challenge, as stated by Douglas and Hurst (1996: 94):

> While it might be possible to evaluate the quality of a document conforming to a specific CL, this does not allow us to say anything about the effects of particular elements in the definition of the CL.

Trying to address this challenge was therefore a key motivation for the undertaking of this study due to the paucity of empirical research conducted on the impact of individual rules.

The undertaking of this study was also spurred by increasing demands for Web-based translated documents. Following a Global Strategies Summit organised by the Localisation Industry Standards Association (LISA), LISA's Chief Analyst, Bill Pullman (2006), admitted that nobody knows the actual size of the opportunities that the localisation industry will be able to generate in years to come. What is certain, however, is that the World Wide Web has dramatically changed the way in which content must be localised to be made available to users. With the emergence of a wide range of new communication channels, such as Rich Site Summary (RSS)

news feeds or alert text messages that are generated from databases, the life-cycle of content has been greatly reduced. This means that the time available to localise content is constantly being reduced. Being able to localise all content is further complicated by the size and growth rates of databases in which content is stored. As far as multilingual communication is concerned, one of the challenges for publishers lies in making their content accessible to most of their global customers, and preferably within a very short timeframe. For many publishers the use of Machine Translation (MT) therefore presents itself as a promising cost-effective solution to address these challenges and produce translated content in timely manner. However, Web authoring guidelines for technical documentation are often not sufficiently precise to ensure the machine-translatability of source content. If stricter rules were used as standards, the performance of MT systems may be improved. The present study will focus on these challenges by examining the impact of specific CL rules on the comprehensibility, usefulness, and acceptability of machine-translated technical documents.

Scope of the study

This study will focus on the reception of machine-translated documents from a Web user's perspective, but will not evaluate the usability of the documents. Usability evaluations focus on the users' actions rather than just on their opinions (Dray, 2004: 31), by studying, for instance, the interaction of users with a given Web interface. However, Web sites now tend to use or reuse chunks of information that originate from databases. This information is then published in multiple formats, using tools that ensure that form is always separated from content. The form of a message is also sometimes referred to as the 'package' (O'Hagan & Ashworth, 2002: 67), dealing with the layout or the selection of fonts and graphics. The final look and feel of a Web document therefore often stems from the template that is used to display the content of databases. Since the purpose of this study is to focus on the content rather than on the form or 'package', a traditional usability study does not seem to be appropriate.

In this study, it was decided to focus on a corpus originating from the IT domain. The corpus contains documents whose main purpose is to instruct, using procedural text. Selecting a corpus from the IT domain is motivated by several reasons. First of all, the CL and MT paradigm may not be new, but it has so far almost exclusively been the preserved territory of the automotive, aeronautic, and telecommunication industries. Apart from Rank Xerox's use of controlled English input to facilitate MT (Adams et al., 1999: 250), a CL and MT project in the IT industry has yet to publish successful results such as those achieved in the automotive sector by companies such as Caterpillar (Kamprath et al., 1998) or Ford (Rychtyckyj, 2006). Times seem to be changing, as SAP was the first company in the software sector to introduce a CL application (Lieske et al., 2002: 1) to ensure the consistency, quality, and machine-translatability of its English and German source documentation (ibid, Schäfer, 2003: 140). Sun Microsystems have also introduced CL in their authoring workflow in 1999, and in 2002 were a year away from introducing MT in their production cycle (Akis & Sisson, 2002: 4). However, no empirical study has yet focused on evaluating the impact of CL rules on the reception of Web-based technical support documentation, despite claims that source quality is a critical factor for the success of such as process (Warburton, 2003, Jaeger, 2004).

This empirical study will be based on a two-phase approach and will fill a gap in the field of CL evaluation, by focusing on the impact of individual rules at the segment and document levels. There is currently no standard methodology available to evaluate the impact of individual CL rules. This study intends to set up an environment that could be reused in future projects to assess the performance of other MT systems with other language pairs, provided that the system uses English as the source language. This study also intends to deliver a clear and unique evaluation of the impact of CL rules on MT output from a Web user's perspective. This will help us determine whether the reception of machine-translated documents can be influenced by the introduction of certain rules.

The structure of the dissertation

The aim of Chapter 1 is twofold. First, it will further explore the context of the present study to discuss the main characteristics of Web-based multilingual technical documentation. This review is performed to identify a text type that is suitable for the present study. Second, it will present the challenges involved in the design, compilation, and description of a corpus corresponding to the selected text type.

In Chapter 2, the concept of CL will be discussed in the light of its use with MT, by reviewing existing CL rules and machine translatability guidelines. The main objective of this review is to select a set of frequent individual MT-oriented CL rules so that the impact that they have on the comprehensibility of MT segments can be evaluated.

Chapter 3 will present the methodology used to extract segments from the corpus and set up an evaluation environment in order to evaluate two sets of MT segments. These MT segments will be obtained by machine-translating two sets of source segments. The first set will contain violations of the rules harvested in Chapter 2, while the second set will contain segments rewritten according to specific CL rules.

Chapter 4 will analyse the results of the first evaluation round, by comparing the scores attributed to the two sets of MT segments created during the setup phase of the evaluation environment described in Chapter 3. The discussion will be based on a comparison of the results obtained for the two target languages. At the end of this chapter, a final list of rules will be drawn and used in the second part of this study.

Chapter 5 will describe the methodology used for the second experimental part of the study, whereby two sets of online documents will be randomly presented to a sample of users of Symantec products. The design of this online experiment and the customer satisfaction questionnaire used will be discussed.

Chapter 6 will present the findings of the second part of the study. The discussion will be based on a comparison of the impact of the rules at the document level, by contrasting results obtained from French and German users.

In Chapter 7, the two objectives outlined earlier will be revisited to determine whether they have been met. The findings of both parts of the study will be summarised and the methodology used will be critiqued. New research questions arising from the findings of this study will be presented to possibly pave the way for future research.

Chapter 1

Chapter 1: Designing a corpus of Web-based technical documentation

1.1. Objectives of the present chapter

This chapter is divided into two sections. Section 1.2 will further explore the context of the present study to discuss the main characteristics of Web-based multilingual technical documentation. The interaction between Web design guidelines and localisation requirements is reviewed in the context of an MT process. This review is performed to identify a text type that is suitable for the present study. Section 1.3 will then present the challenges involved in the design, compilation, and description of a corpus corresponding to the selected text type.

1.2. Characteristics of multilingual Web-based technical documentation

1.2.1. Web localisation bottleneck

The area of Web localisation is sometimes perceived as the 'fastest-growing area in the translation sector today' (O'Hagan & Ashworth 2002: ix). This is no surprise considering the ever increasing amount of content to be translated in a very limited period of time. Even though localisation does not only involve translation, publishers are often striving for a simultaneous publication of their information in multiple languages. As far as multilingual Web sites are concerned, Esselink (2001: 17) also warns that 'the frequency of updates has raised the challenge of keeping all language versions in sync (...), requiring an extremely quick turnaround time for translations.' However, providing information before it becomes obsolete is sometimes not possible for publishers, and some content is published exclusively in the language in which it has been authored. Yunker (2003: 75) remarks that 'unless the target audience consists of only bilinguals, this approach is bound to leave people feeling left out.' This is confirmed by O'Hagan & Ashworth (2002: 52) who point out that 'readers require a given Web page to be in their language to allow for real-time browsing and information gathering'. In 2000, the International Data Corporation carried out a survey within the framework of the Atlas II project. Based on the results obtained from 29,000 Web users, they estimated that by 2003, 50 per cent of Web users in Europe would be likely to favour sites in their native language

(Myerson 2001:14). This trend is noteworthy, as the number of Internet users that are non-native English speakers grows by over 140 million per year (Levin, 2005: 45). The lack of global distribution and accessibility has been highlighted by Pym (2004: 9) and is reflected by three types of locales:

- The 'participative' locale consists of users who are able to access information in a language they can understand. These users are then able to act upon the information they have accessed.

- The 'observational' locale consists of users for whom it is too late to do anything with the information they access. They are able to access it in their own language, but by the time this information is translated, it is obsolete.

- The 'excluded' locale consists of users who are never given the chance to gain access to information in a language they can understand.

For certain users, this accessibility problem may be alleviated by the use of online MT services provided by portals such as Google or AltaVista Babel Fish. If excluded users cannot get access to information in their own language, they may resort to online MT services to attempt to understand foreign language information.

1.2.2. Free MT services or refined MT services?

The use of MT for the translation of Web content has been described by Hutchins (2004: 16):

> Closely related to the use of MT for translating texts for assimilation purposes is their use for aiding bilingual (or cross-language) communication and for searching, accessing and understanding foreign language information from databases and webpages.

With this new approach, users of free online MT services may be able to get the 'gist' of a Web page thanks to raw MT output. 'Raw MT output' is used to refer to the translation produced by a general purpose MT system, which was not customised in any way for its end-users. Informal reports suggest that these users show a pragmatic attitude towards the capabilities of these systems (Somers, 2003: 523), which may help them overcome the barriers of a communication deadlock. However, such a practice 'goes against the general recommendations for the use of

13

MT that have been made over the last decade or so' (ibid). As indicated by Yang & Lange (2003: 200), it is difficult to evaluate the usefulness of MT output for these users. This seems especially true when the type of content published is domain-specific and requires a certain level of quality. For instance, in the case of online technical documentation, users must be able to act on the information they are given to read. A general purpose MT system whose dictionaries do not contain specific terminology will probably prove inadequate in these situations. In order to try and address this issue in a cost-effective manner, certain software publishers, such as IBM, Cisco, and Microsoft have therefore started providing online refined MT content to the users of their online knowledge bases (Warburton, 2003, Jaeger, 2004, Richardson, 2004).

Uncertainties, however, remain as to whether refined MT output can be accepted and used effectively as a translation product by Web users, especially when source content is not controlled. Controlling source content in a Web localisation context may be regarded as a Web globalisation task, since its objective is to ensure that as many locales as possible can gain *access* to a certain type of information. Section 1.2.3 will review some of the guidelines used for the development of Web content to determine the extent to which they contribute to making content more accessible and translatable in a Web globalisation context.

1.2.3. Web globalisation

According to Pym's terminology (2004: 9), the localisation process attempts to transform 'excluded' locales into 'participative' locales rather than 'observational' locales. Since this process might involve the publication of information in multiple languages, publishers must plan ahead to ensure they can quickly cater for all their multilingual customers. This challenge is commonly regarded as a Web globalisation issue, whereby Web content should be designed and maintained with localisation in mind (Esselink, 2003: 68). From a translation perspective, efforts are being made by working groups within the World Wide Web Consortium (W3C) to standardise the handling of specific content within XML documents, so as to simplify the localisation process of Web content. The W3C is the organisation that works on creating and maintaining Web standards. For instance, the Internationalisation Tag Set (ITS) provides translatability rules so that translators

know whether an element must be translated or not'[1]. This approach seems particularly adapted to technical information, which often contains technical words or chunks of text that should not be parsed by a general MT system. However, non-translatable content may not be systematically marked-up. This issue is raised by Schachtl (1996: 145), who states that 'a common fault in software documentation is the insufficient marking of chunks of object code in running text.' The metadata-based approach of the ITS also proposes the possible tagging of ambiguous terms to ease the translation process. From an MT perspective, the challenge lies in ensuring that MT systems can understand these rules and act on them. Whereas the ITS focuses on the translatability and localisability of Web content, specific Web design guidelines deal with the actual accessibility of the content. These guidelines are reviewed in the next two sections.

1.2.3.1. Web accessibility guidelines

In his foreword to *Maximum Accessibility: Making your Web Site More Usable for Everyone* (Slatin and Rush: 2003), Nielsen states that Web usability and Web accessibility are 'two tightly intertwined concepts', because content that is made accessible to a certain group of users 'tends to be usable for all users'. Existing linguistic Web accessibility guidelines (W3C, 1999) focus on simplicity and clarity, so it may be hypothesised that they share certain features with CL rules. In this light, violations of CL rules may impact Web accessibility. This use of the term 'accessible' does not correspond to the strict definition of 'Web accessibility' used by Slatin & Rush (2003: 3). For them, Web sites are deemed '*accessible* when individuals with disabilities can access and use them as effectively as people who don't have disabilities'. This strict use of the term 'accessible' refers to section 508 of the US Rehabilitation Act of 1973, which was amended by Congress in 1998. However, the semantic scope of this term has since widened to encompass groups of users that contain more than just users with physical disabilities. For instance, Nielsen (2000: 309) mentions that cognitive disabilities should not be underestimated.

---

[1] For more information, see http://www.w3.org/TR/its/#translate [Last accessed: September, 18th 2006]

15

It may even be argued that people who are not competent in a certain language can be affected in the same way as, say, visually-impaired people. The W3C has identified these new challenges, and use the term 'Web accessibility' to refer to a wide range of situations. In their guidelines, they suggest that 'many users may be operating in contexts very different' from the situations that Web developers are familiar with. Certain users may not be able to perceive or interact with Web content, which prevents them from having access to the information they need. The W3C (1999) draws the following list of users who may find certain Web content non-accessible:

1. 'They may have a text-only screen, a small screen, or a slow Internet connection.

2. They may have an early version of a browser, a different browser entirely, a voice browser, or a different operating system.

3. They may not have or be able to use a keyboard or mouse.

4. They may be in a situation where their eyes, ears, or hands are busy or interfered with.

5. They may have difficulty reading or comprehending text.

6. They may not speak or understand fluently the language in which the document is written'.

This list shows that challenges to Web accessibility can originate from a wide range of issues. They can be of a technical nature, as in situations 1 and 2, of a physical nature as in situations 3, 4, and 5, and of a cognitive nature as in situations 5 and 6. Technical and physical issues, which mainly pertain to the form of Web content, are abundantly addressed in the W3C guidelines and specialised literature. However, cognitive issues, which are often linked with linguistic accessibility, seem to be neglected. If we were to use Klare's definition of accessibility (1963), we could say that the 'ease of understanding or comprehension due to the style of writing' is absent in non-accessible Web sites. This is valid for Web content which is not localised, and therefore non-accessible for the group of users who are not competent in the language in which the source content has been published. But to some extent, this also applies to Web content which is too difficult to comprehend, even for certain native speakers, due to their lack of technical expertise. Some users may

then be able to *physically* and *technically* access specific Web content without being able to get the answer they may be looking for. On the other hand, people with the proper technical background may not be able to *physically* or *technically* gain access to Web content they would able to understand. Specific writing guidelines used to make Web content more accessible and comprehensible are discussed in the next section.

1.2.3.2. Writing guidelines for Web content

Linguistic Web accessibility guidelines are not always the main concern of Web accessibility experts. In their preface to *Maximum Accessibility: Making your Web Site More Usable for Everyone,* Slatin and Rush (2003: xxiii) inform readers that they will tell them 'how to make the World Wide Web more accessible and more usable for everyone around the world'. In this same preface, they mention the W3C's 14th guideline concerning the clarity and the simplicity of the documents. Yet, they do not refer to this guideline anywhere else in their book.

More detailed writing guidelines are presented by Spyridakis (2000: 376) to produce comprehensible Web content. She mentions that certain text features will impact the reading process:

> If readers can devote less attention to lower level tasks such as decoding letters, words, and syntactic structures, they will have more attention available for higher level tasks, such as combining text-based information with other text-based information and also with information stored in long term memory.

In order to avoid such processing issues, Spyridakis provides a list of guidelines that are destined to improve the comprehensibility and translatability of Web content[2]. However, most of these guidelines are not precise enough to be implemented on a systematic basis, an example being the use of 'simple sentence structures and internationalized words and phrases'. This lack of precision in the definition of specific guidelines was also found in *Designing Web Usability: the Practice of Simplicity* (Nielsen, 2000: 101). Nielsen provides general advice without getting into great detail. He mentions that the three main guidelines for writing for the Web

---

[2] These guidelines are available at:
http://www.uwtc.washington.edu/research/pubs/jspyridakis/Quicklist_Comprehensible_Web_Pages
[Last accessed on August, 25th 2006]

concern text conciseness, content scannability, and the presence of hyperlinks. The first two guidelines share one common point: they are difficult to apply on a systematic basis due to their general nature. They do not make explicit the means which can be used to render content clearer and simpler. These guidelines' lack of precision does not benefit users of technical documentation.

The first guideline, conciseness, echoes a general principle of technical communication, whereby writers are often asked to avoid verbosity (D'Agenais and Carruthers, 1985: 100; Gerson and Gerson, 2000: 31; Raman and Sharma, 2004: 187). However, excessive conciseness may sometimes have a negative impact on the clarity of messages, due to ambiguities introduced by the compression of information (Byrne, 2004: 24). By removing words such as articles or prepositions from the source text, the clarity of a message may be affected. This issue is likely to occur if guidelines are too stringent. For instance, Gerson and Gerson (2000: 243) suggest that Web content sentences should contain between 10 and 12 words. If this guideline were to be enforced in a systematic manner, essential syntactic components would be removed from certain sentences.

Next, the scannability of Web content seems motivated by two factors. First, it is believed that users take 20-30% longer to read from a screen than from a page, and second, they rarely read a source text in its entirety (Nielsen, 2000: 104). They tend to focus on the part of the text that most interests them. Pym (2004: 187) even goes as far as saying that users are 'no longer readers' due to the loss of discursive linearity of texts.

This review of Web content and Web accessibility guidelines suggests that precise rules are required if MT systems are to be used effectively for the production of useful Web-based translated information. As briefly discussed in the introduction, certain CL rules may have the potential to complement these guidelines in order to improve the machine-translatability of Web-based technical documentation. Before such rules can be proposed as standards, their effectiveness must be evaluated at the segment level and document level. In order to achieve these objectives, test content is required. Section 1.3 discusses the way in which test content will be assembled, by examining the challenges involved in corpus design, compilation, and description.

1.3. Corpus design, compilation, and description

1.3.1. Impact of the study's objectives on corpus design requirements

Kennedy (1998: 70) remarks that the 'optimal design of a corpus is highly dependent on the purpose for which it is intended to be used'. The objectives of this study must therefore be assessed to determine their impact on corpus design. In this study, the corpus will be used to achieve the following objectives in a two-phase approach:

- to provide multiple violations of CL rules to evaluate the impact of these rules on the comprehensibility of refined MT segments

- to provide genuine technical documents that will be machine translated and published within the framework of an online experiment involving users who require localised information

A specific corpus must be identified in the field of Web-based technical documentation so that these two objectives can be achieved.

The two objectives have implications with regard to the audience of the text type used in this study. Technical documentation may be produced for a wide range of users, so the level of technical information will vary depending on the audience targeted. In the IT domain, two main groups of users can be identified based on the distinction that is often made between home and enterprise software products:

- Home users, who may or may not have any technical knowledge

- Enterprise users, such as system or network administrators, who should have an excellent technical background

Even though the difference is not always easy to ascertain, most technical documentation targets at least one of those categories of users. This difference, however, has an influence on the communicative function and localisation requirements of a given text type.

This separation between general and professional users impacts on the features of the speech act that will be used in the communication process. Sager (1994: 52) uses the term *speech act* 'to designate functionally coherent interactions performed by means of linguistic signals'. In short, the linguistic features of any published

technical message will be based on the type of interlocutors involved in the speech act. Technical content developers are specialists, whereas readers or users will either belong to a lay person category or a specialist category. The dichotomy involved with the participants of the speech act will therefore impact on the linguistic materialisation of the technical message. According to Sager (1994: 45), the language used between specialists will hover between 'artificial language' and 'special subject language' used in an overall 'specialised discourse'. On the other hand, the language used between specialists and lay-persons will only use 'artificial language elements' used in an overall 'general discourse and popularised specialist discourse'. Since the text type chosen for this study must not be too specific, it should be targeted at home users, who are more likely to require translated documentation than their enterprise counterparts.

The second objective also has technical implications. As MT content will have to be published in a real-life situation, the repository of the chosen text type should allow for the storage of localised content. Kennedy (ibid) also insists on three central issues in corpus design: whether the corpus should be static or dynamic, the representativeness of the corpus, and its size. These issues will be reviewed in the light of our objectives so as to obtain a final list of corpus design requirements.

1.3.1.1. Static or dynamic corpus?

The main objective of this study is not to monitor the evolution of a certain technical authoring style over time. Rather, raw textual data is required for the first part of this study, so that the impact of CL rules on the comprehensibility of refined MT output can be evaluated at the segment level. Olohan (2004: 45) maintains that a static corpus provides a 'snapshot of aspects of the language at a particular point in time.' This does not mean, however, that the focus should only be placed on content that has been written in a limited period of time.

If violations of existing CL rules are to be found in the corpus, the content should ideally have been authored over a number of years and still be accessed by users.

So far the terms 'content' and 'documents' have been used without providing any strict definitions. Traditionally, technical documents (or articles) are composed of

content which can be updated on an irregular basis (only when new information is available), or on a regular basis (to match the quarterly or yearly release of a new product version). So a technical document may well have been *physically* created a few years ago, but some of its content may have slightly or radically altered over time. According to Hammerich and Harrison (2002: 2), the term 'content' refers to the 'written material on a Web site', whereas the 'visuals refer to all forms of design and graphics'. The selected corpus should therefore integrate any content that could be machine-translated in the second part of the study, even if it was written a few years ago. Some of the linguistic constructs used in the content developed a few years ago may no longer appear in new documents. For instance, certain content developers may have stopped using certain syntactic patterns or specific phrases. However, these patterns may still be present in older technical documents, so they should not be omitted from the evaluation in the first part of the study. This legacy content may not be as valuable as new content, but it is worth investigating because it may still be relevant for certain users.

Technical content is subject to frequent updates. Thus, it was anticipated in the domain of this study that there would be a discrepancy between the textual content of the documents gathered for the design of the corpus, and the textual content to be machine-translated in the second part of the study. By the start of the second part of this study, some of the documents may have been updated. Since the overall objective of this study was to use real content for the evaluation of CL rules, the updated versions of the documents were used.

1.3.1.2. Representativeness of the corpus

One of the objectives of this study is to focus on the effectiveness of individual CL rules. In order to be able to evaluate their impact effectively, violations of these rules must be found in the corpus.

If the selected text type is too specific, certain CL rules may not be violated. The text type chosen for the design of the corpus should therefore not be too specific so that genuine rule violations can be found. As mentioned by Olohan (2004: 45), the representativeness of the corpus affects the 'degree of conviction' when it comes to making generalisations based on the data analysis. It is therefore essential to ensure that:

- The chosen text type is not specific to Symantec or the Internet security industry. The text type should be common in the software industry

- The syntactic and textual features of the content should not be considered as sublanguage characteristics. They should ideally be echoed in other types of documentation so as to obtain a balanced corpus

- The content has been authored by a range of developers, who may or may not be native speakers of English. This will helps us obtain a varied list of CL rule violations

- The content to be included in the corpus should contain technical information destined for home users who need localised material

- The content should be published via the Web

- The underlying content repository should already contain localised materials to ensure the seamless storage of future machine-translated content

These considerations are essential to take into account to ensure that the present study acts as a genuine case study. If certain CL rules prove effective, they may be reused in the future for similar text types in the industry. Lalaude et al. (1996: 112) state that certain rules are sometimes document-specific, while others do tend to be generic and may be applied to different types of documents.

1.3.1.3. Size of the corpus

The size of the corpus should reflect the linguistic diversity requirement, so that a wide range of violations of CL rules can be identified. However, the corpus must not be extremely large for the following reasons. Firstly, it can be time-consuming, even for a specialised piece of software, to process a large corpus.

The experimental approach used to extract examples from the corpus will be described in greater detail in Chapter 3. But it is essential to take into account hardware resources when designing the corpus. The larger the corpus, the longer it takes to perform an analysis of some sort, especially when the objective is to refine that analysis over time. For that reason, a large corpus in the region of a million words had to be discounted.

Moreover, the second objective of this study is to evaluate the impact of certain CL rules within an online experiment using machine-translated material. In order to publish this material, source content can be manually edited so as to conform to the CL rules identified in the first part of this study. Due to time constraints, only a limited number of documents will have to be edited prior to their machine translation. It is therefore necessary to have enough documents to choose from, but not essential to have an exhaustive set of documents.

Finally, adding more documents or parts of documents to a corpus when those documents belong to the same text type and domain can lead to a 'content bottle-neck'. Even though the word count and the size of the corpus increase, there is little new data. Bowker & Pearson (2002: 48) therefore write that 'it is generally accepted that corpora intended for Language for Specific Purpose studies can be smaller than those used for Language for General Purpose studies (..). In our experience, well-designed corpora that are anywhere from about ten thousand to several hundreds of thousands of words in size have proved to be exceptionally useful in LSP studies'. Based on these guidelines, it was decided to look for a text type that would be able to produce a corpus size of approximately 200,000 words.

1.3.2. Selecting the ideal text type for a case study

At the start of the study, in 2004, Symantec's technical content focused predominantly on security applications. As some documents may be semantically related, but have different communicative purposes, and in turn different syntactic distributions, it was essential to review several text types to see whether they met the corpus requirements. This quick review was essential to select a text type that was not specific to Symantec or to the Internet security industry, but rather, common in the software industry. If the text type were specific to Symantec, it would be difficult to conduct a relevant case study. For instance, technical articles focusing on the description of security threats are specific to the Internet security industry. Besides, the repetitiveness of their content would make it difficult to find genuine examples of violations of CL rules. This text type uses a language that can be described as a sublanguage due to the closed nature of the syntactic structures being used and the large amount of non-translatable elements. Since the objective of

this study is to evaluate the effectiveness of as many CL rules as possible, a more generic text type was required. Special attention was paid to technical support documentation since this type of text is very common in the software industry.

1.3.2.1. Technical support documentation

Web-based technical support documentation is often used by software companies as an after-sales service provided to users experiencing technical problems with a particular software application. Marlow (1995: 44) states that support documents 'provide additional task-related information not otherwise covered in the user manual to help users solve particular problems'. The purpose of these documents is to reduce customer contacts by offering online technical content so that global customers can find answers to their questions or resolve their problems by themselves. According to Freeman (2006: 4), this type of content is 'cost-reducing content' because it helps reduce customer service costs. From a financial perspective, users may also benefit from this type of content, especially if they have to pay for support services. Marlow (2005: 5) believes that 'users would prefer not to contact technical support technicians, and this is especially true when support systems are chargeable'. However, specific information pertaining to Web-based technical support documentation is scarce in the literature on technical documentation, which often focuses on user's guides and online help systems (Price and Korman: 1993; Marlow: 1995; Gerson and Gerson: 2000; Raman and Sharma: 2004). This type of information is equally scarce in the literature related to online content development (Spyridakis: 2000; Hammerich and Harrison: 2002).

In a traditional model, these technical support documents are stored in knowledge bases, which are special databases for knowledge management using a well-defined classification structure. The documents are then published via the Web so that they can be easily accessed by users. Users have then two ways to access these documents. They may follow a link present in some of the software application's dialog boxes so as to be redirected to an online document. They can also go to the company's Web site and search the knowledge base online. The method used will depend on the type of problem encountered by users. Let us take the example of a user experiencing an error message when trying to install a product. Such an error message may be due to a genuine bug in the application, or to a mistake made by

the user during the installation process. In the second scenario, a document explaining how to avoid this issue may have been published online if the issue was experienced by other customers. The error message's dialog box seen by the user may then contain a hyperlink referring to this online technical document.

Certain documents are also authored to show users how to perform common tasks with their applications, rather than to fix a technical problem. From a communication perspective, the authors of enterprise technical support documents address specialists, whereas consumer technical support documents provide an environment where specialists advise lay persons. Even though certain home users may be computer literate, it is most likely that people who need to read those documents are people who cannot address the initial problem. The importance of localised consumer material is therefore crucial to ensure the accessibility of the content to all of the users. It is hypothesised that most home users with a basic knowledge of English would not be able to solve a problem by reading instructions in a language they barely understand. As most of technical support content is prescriptive, missing or misinterpreting a step can have negative consequences. This fact alone seems to justify the translation of consumer technical support content into various target languages. Regardless of the reasons for not fully localising Web content (time, cost, lack of resources), the consequences of having Web content that is only partly localised should not be underestimated. The absence of Web localisation can be compared to the absence of localised technical documentation. Hoft (1995: 205) believes that 'non-native readers of the source language can become dissatisfied with a technical manual in the source language and project this dissatisfaction onto the company.'

This challenge has also been pointed out by Van der Meer (2003: 181):

> It is no longer sufficient to translate the product documentation and localise the software, all enterprise information must be adapted to the global marketplace in which the company wants to be seen as market leader. (...) Current users will buy online, get support and get training online.

As aforementioned, translating every published technical support document is a challenge for large software companies, including Symantec, due to the size of their

knowledge bases. The amount of diverse content they contain appears a good starting place to design a corpus.

As far as this study is concerned, it is important that the linguistic coverage of this type of content corresponds to the representativeness criterion which was defined earlier. It may be argued that Symantec consumer technical support content only applies to Symantec consumer products and users. As Symantec consumer products are mostly Windows-based, the domain is also reduced. But this study focuses on the types of linguistic structures used in a particular text type rather than on its lexical composition. The findings, which will be made based on a particular sample of Symantec-specific consumer technical documentation, may also apply to other technical support documents within the software industry. Every software publishing company provides technical support to their customers, regardless of the type of product they manufacture. It is unlikely that a slightly different semantic field would undermine the findings of this study based on a Symantec-specific corpus. Of course, these claims must be justified in greater detail by examining Symantec consumer technical support documentation. This will help us produce a macro-level description of this text type and determine whether certain characteristics of technical support documentation can be found in other text types. The next section focuses on the textual characteristics of this text type from a user's point of view.

1.3.2.2. Characteristics of a technical support document: the user's perspective

Based on the classification of macro text types in science and technology given by Sager (1994: 85), technical support documents seem to fit in the 'memo' category, because they are 'concerned with the immediate interaction of participants and topics'. Users need to consult technical support documentation when they have technical problems or questions that remain unanswered. As mentioned previously, they may get direct access to a particular article by clicking on a hyperlink from within their application. They may also get access to a solution by going online and searching the technical support's knowledge base. In the second scenario, they can view the most frequently accessed documents after selecting search criteria such as product name and version name. End-users may also enter a keyword corresponding

to their query to retrieve a list of search hits, from which they can select the document that best corresponds to their needs. A typical technical support document as viewed in a Web browser is shown in Figure 1-1 below.
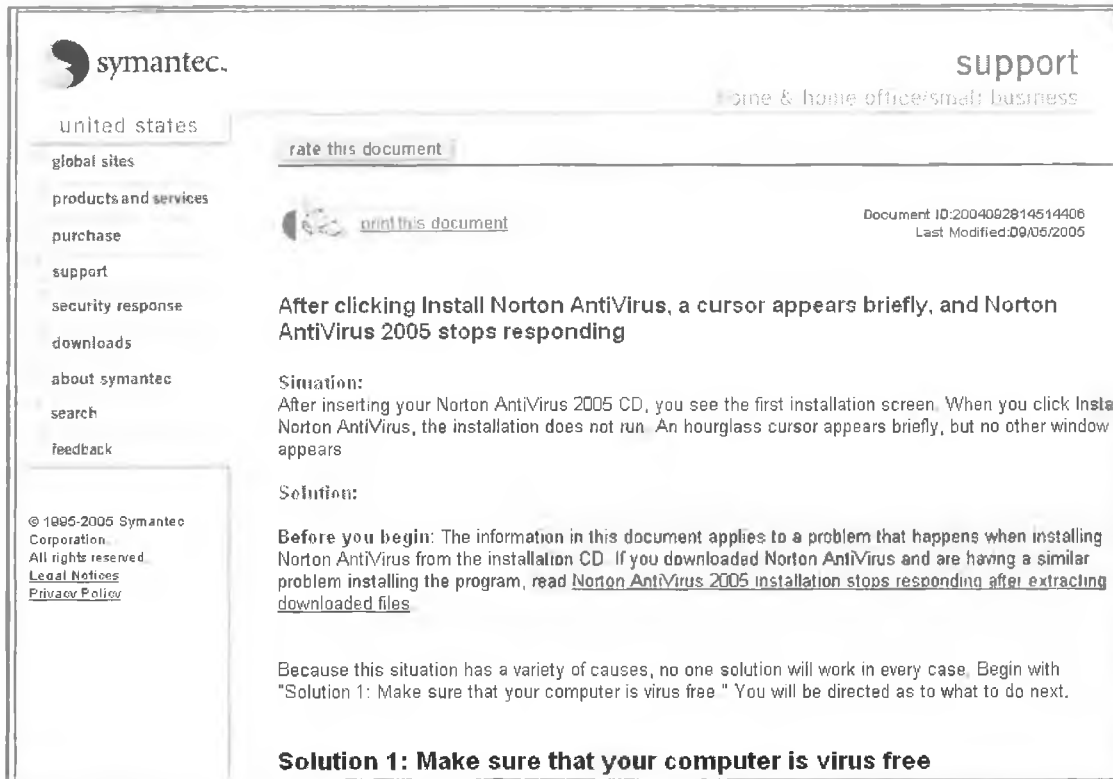
Figure 1-1: Technical support document viewed in a Web browser

A typical Web-based Symantec consumer technical support document is a document that can be displayed in a traditional Web browser. If a text-only Web browser is used, the document may not be complete because of the possible presence of non-text elements such as graphics.

1.3.2.2.1. Linguistic elements of graphics

Some of the graphics present in technical support documents are 'screenshots', showing specific parts of an environment in which something happens or needs to be done. The term 'environment' refers here to the Graphical User Interface (GUI) of a program or a set of programs. The use of screenshots establishes a strong connection between technical support documents and software user guides, the latter containing sections that show users how to utilise a particular piece of software. Byrne (2004:3) defines user guides as 'the interface between computer systems and human users'. In user guides, sections are often illustrated with

screenshots whose purpose is to guide users in step-by-step situations, such as activating a particular function, modifying certain settings, or removing an application. Since screenshots sometimes perform the same function as text instructions, one may wonder why one is used instead of the other, or why both are sometimes used together. Elements of answers to this question may be found in a study (Fukuoka et al., 1999) which found that American and Japanese users believe that more graphics, rather than fewer, make instructions easier to follow. This study also revealed that users prefer a combination of text and graphics, which they believe would be more effective than text-only instructions.

From a semiotic perspective, screenshots play an iconic role (Dirven & Verspoor, 1998: 4), because they provide users with a replication of the environment with which they are interacting. Screenshots may also provide an illustration of some of the steps users should follow to fix a problem. Technical support screenshots can sometimes be edited by content developers to provide extra information to users. Information can be added using text or graphical drawings, such as arrows or circles, to draw the attention of the user to a certain part of the replicated GUI. This is exemplified by Figure 1-2:

Windows 95/98/Me/NT/2000

1. Cliquez sur **Démarrer** > **Rechercher ou Trouver** > **Fichiers ou dossiers**.
2. Assurez-vous que (C:) soit bien sélectionné dans la boîte de dialogue Rechercher dans et que l'option **Inclure les sous-dossiers** ait également été choisie.
3. Dans le champ "Nommé" ou "Rechercher...", saisissez -ou copiez/collez- les noms de fichiers suivants :

downloads

4. Cliquez sur **Trouver maintenant** ou **Rechercher maintenant**. Les résultats sont affichés dans le volet de droite ou dans la partie inférieure de la fenêtre. Il peut y en avoir plus d'un. Le dossier que vous recherchez est celui dont le chemin d'accès finit par \LiveUpdate.

Name	In Folder
Downloads	C:\Documents and Settings\Administrator\My Documents
Downloads	C:\Documents and Settings\All Users\Application Data\Symantec\LiveUpdate
downloads.gif	C:\Program Files\Ationline\DOC
Downloads	C:\Documents and Settings\Administrator\Recent

5. Cliquez deux fois sur ce dossier.
6. Supprimez tous les fichiers ou dossiers contenus dans le dossier \LiveUpdate\Downloads. Dans la plupart des cas, il s'agira de Autoupdt.trg et Livetri.zip. Un ou plusieurs dossiers peuvent également s'y trouver. Supprimez-les tous.
7. Exécutez LiveUpdate de nouveau. Si le problème persiste, passez à la section intitulée "Un fichier Livetri.zip corrompu".

Figure 1-2: Edited screenshot

These elements are examples of an indexing principle (Dirven & Verspoor, 1998: 5) because they draw the attention of the user to a particular action that should be performed, or to the result of an action. This allows users to isolate the component of the GUI which requires action. As a result of this quick link between form and meaning, screenshots may replace procedural sentences containing instructions to find the location of a graphical item, be it a button, a tab, a pane, a window, a menu bar, a menu item, or a radio button. These elements may also have an iconic function by replacing the action that the user should perform on one of these items: to click, to check, to uncheck, or to enter a word.

From a multilingual communicative perspective, those screenshots should of course be in the language of the users so that their primary iconic function can be fully performed. However, this is not always possible, because third-party English applications are not always localised. A document that is written in French but which contains non-localised screenshots, where the language used in the UI is English, is therefore bound to baffle French users, as shown by Figure 1-3:

Figure 1-3: Non-localised screenshot in a localised document

The handling of screenshots is therefore a complex localisation process, which is beyond the capabilities of an MT system. As shown by the previous example, it is sometimes difficult or impossible for a human translator to find the corresponding screenshot in his or her own language. The time required to perform such a search during the translation process should therefore not be underestimated. This is not the only drawback of screenshots when they are included in technical support documents. Screenshots can also create accessibility issues for users with eyesight-related disabilities. If screenshots are not accompanied by alternative text, they may be ignored by accessibility tools such as screen narrators. Figures 1-4 and 1-5 show an excerpt of a technical support document with a screenshot, and the underlying HTML code from which the alternative text string is missing:

4. Klicken Sie im Registrierungseditor auf den Schlüssel HKEY_LOCAL_MACHINE.
5. Klicken Sie im Menü **Sicherheit** auf **Berechtigunge**.

Berechtigungen für HKEY_LOCAL_MACHINE

Sicherheitseinstellungen

Name	
Administratoren (ADMIN-HATSM7FEW\A...	Hinzufügen...
EINGESCHRÄNKTER ZUGRIFF	Entfernen
Jeder	
SYSTEM	

Berechtigungen: Zulassen   Verweigern

	Zulassen	Verweigern
Lesen	☑	☐
Vollzugriff	☑	☐

Erweitert...

☑ Vererbbare übergeordnete Berechtigungen übernehmen

OK   Abbrechen   Übernehmen
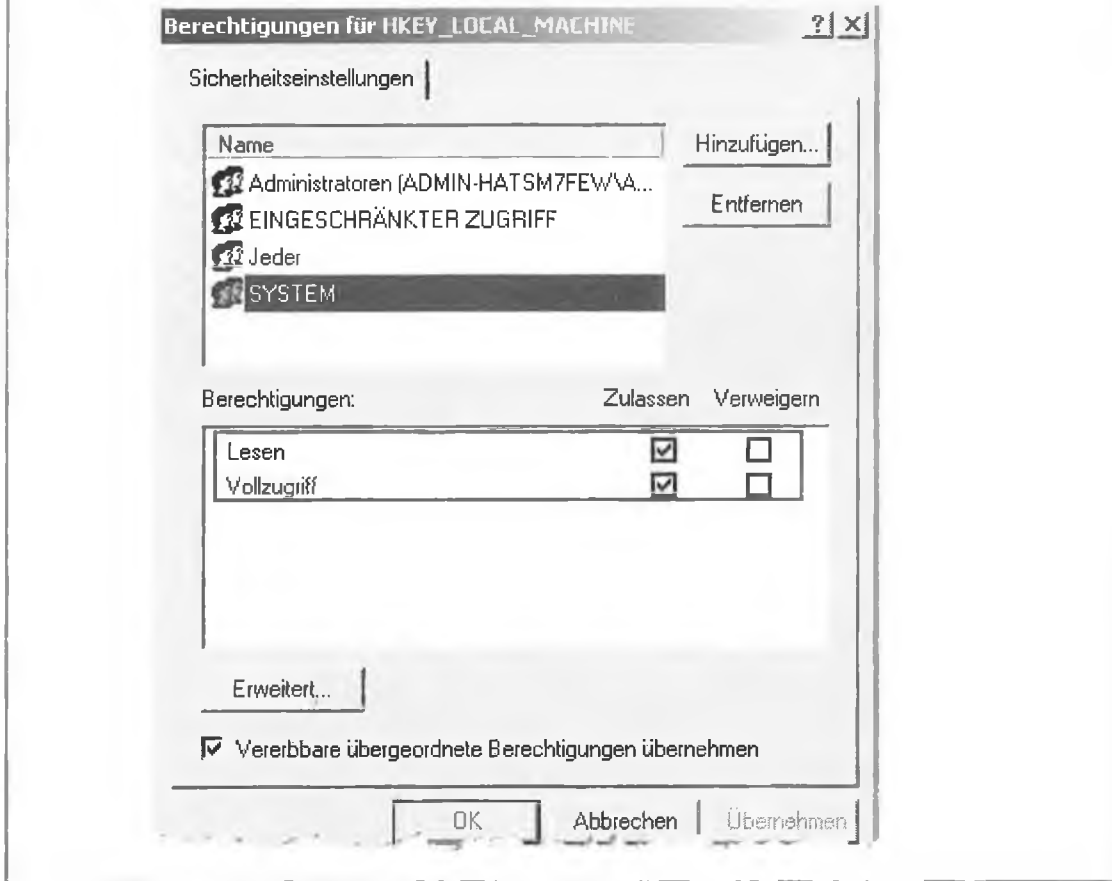
Figure 1-4: Localised screenshot in a technical support document

```
<li>Klicken Sie im Registrierungseditor auf den Schlüssel HKEY_LOCAL_MACHINE.
<li>Klicken Sie im Menü <b>Sicherheit</b> auf <b>Berechtigunge</b>. <br>
<br>
<img
src="/SUPPORT/INTER/navintl.nsf/bf1fb726f6e7892985256ef50048caf2/0d312f63245654a580256d6400396e
d2/Solution/0.1EC8?OpenElement&amp;FieldElemFormat=gif" width="370" height="439" alt=""><br>
<br>
<br>
```

Figure 1-5: Missing alternative text in HTML code

Besides, a screenshot may impact on the reliability of a document over time, or at least baffle users running older versions of the product for which the document was originally intended. The iconic function of screenshots was mentioned because of the link it created with a given GUI at a certain point in time. However, that link may well be altered or even broken if the GUI changed over time. For instance, a document applies to several versions of an operating system when the text used in the document does not focus on any particular version. If a screenshot is introduced, the document may be perceived as version-specific by certain users. If the

screenshot does not exactly match their environment, certain users may be tempted to think that the document does not apply to them.

To some extent, this also applies to video clips that are sometimes linked to technical support documents. These video clips contain step-by-step tutorials designed to help users find an answer to their question. From a localisation perspective, this type of element is even more complex than static screenshots, while having the same pragmatic function as plain text. Plain text is, of course, the most common element among the textual elements that are contained in technical support documents.

1.3.2.2.2. Textual content

The screenshot in Figure 1-1 outlined the typical structure of Symantec's technical support documents, which are based on a template. Certain sections are compulsory, but the text they contain is not subject to any control[3]. All documents consist of a title, which summarises the technical issue or question encountered by the user, and of a body of text containing several sections. Titles are directly followed by a 'Situation' section which expands on the information present in the title. The 'Situation' section establishes an immediate one-to-one relationship between the writer and the user, as shown by the frequent use of the second person personal pronoun 'You'. By reading this section, users should then be able to determine whether they are reading the right document. The situation that is described in these first introductory lines should be *exactly* the same as the one they were experiencing before accessing the document. The 'Situation' section is pivotal to the resolution of users' problems, because the subsequent sections of the document are based on the assumption that users are experiencing a specific problem in a specific context with a specific feature of a specific application.

---

[3] In 2005, Symantec decided to deploy an application with CL checking capabilities (acrocheck ™) to improve the consistency, readability, and machine-translatability of its new technical support documents. The selection of the corpus was carried out prior to this deployment.

The 'Solution' section is also always present in Symantec's technical support documents. It contains all the information necessary to answer a question, or solve a problem. As shown in Figure 1-2, it may contain a 'Before you begin' segment to remind users that the solutions described in the section relate to a specific problem. The rest of the 'Solution' section contains a list of steps that users must follow to resolve their problems. The nature of this section is procedural since the steps users have to follow are often presented using numbered lists. Price and Korman (1993: 227) state that the aim of such a procedure is 'to carry out a small-scale task'. This is achieved by ensuring that 'each step specifies an action readers take on the way toward accomplishing their goal' (ibid: 233). Most of the list items therefore start with imperative verbs. In certain cases, series of Prepositional Phrases (PP) are used before the imperative verb, as in: 'In the main window, on the left side, click Norton AntiSpam'.

Symantec's technical support documents may also contain two other types of sections: a 'Technical Information' section and a 'Reference' section. These descriptive sections are not always essential for solving the issue discussed in the document. Instead, they provide a starting point for users who want to learn how to avoid future issues. For instance, the 'reference' section may point to internal technical documents within the same knowledge base, but it may also refer to third-party material. These links may be seen as instances of intertextuality because content developers assume that a certain technical issue will become clearer if users become familiar with the subject field. The next section of this chapter focuses on the textual elements that facilitate intertextuality: hyperlinks.

1.3.2.2.3. Intertextuality: the role of hyperlinks

A typical textual feature of technical support documents lies in their use of hyperlinks to refer users to background articles or to related articles. Those related articles may prove invaluable in the resolution of a problem, especially if users realise that they have been looking at the wrong document. Hyperlinks deconstruct the traditional monolithic structure of a document, by offering users the possibility to get access to information from several documents. As long as users can find all the information they require, it does not matter whether they have read some or all of one or more documents. What is, however, crucial in this approach, is to have

hyperlinks referring to target content which is in the same language as the original content. Legitimate frustration can sometimes emanate when the localisation chain is broken in the middle of a quest for information. This aspect of technical documents will be borne in mind in the second part of our study, to ensure that published documents can either stand on their own or refer to localised documents. Now that the main textual characteristics of technical support documents have been highlighted from the user's point of view, it is possible to proceed to a detailed review of the content from the developer's perspective.

1.3.2.3. Characteristics of technical support content: the developer's perspective

1.3.2.3.1. Technical support knowledge base

Thus far, the focus has been placed on technical support documents from a user's perspective, with little focus on the situations that arise when users need to consult a technical support document. End users' typical needs can be analysed by examining the classification used to store technical support articles in Symantec's knowledge base. Sager (1994: 101) states that the exchange of knowledge is 'the basic condition for success in non-phatic communication'. When developers add new content to the knowledge base, they must have good reason to do so, which is often to bridge the knowledge gap between themselves and the users. The categories used by developers to label documents are therefore a good indicator of what is usually requested by the users.

1.3.2.3.2. Classification of knowledge

When a document provides a solution for users experiencing a general technical problem, the article is classified into sub-categories based on the issue encountered: whether it is a compatibility issue, a third-party issue, or a functionality issue. The particular feature of the application in which the problem occurs can also be used as a filter. A document can thus be classified in several categories, so as to maximise chances of retrieval when the information is required. Such documents are written to provide an answer to an open question introduced by an interrogative pronoun such as 'why', 'how', 'what', 'which' or 'where'. According to Price and Korman (1993: 320), such documents should be limited to one 'unit of information which is the topic'. For instance, a user may not be able to use a specific feature of the

program and would like to know the reasons why. This type of document provides an answer a posteriori to help users resolve a problem that has already occurred. Specific technical problems are often classified in an 'Error Message' category, because the error message number allows for quick and easy identification of the problem.

In other cases, users may be looking for more information prior to experiencing a problem. They may have searched to no avail in the online help or in the user manual. In this scenario, the type of question asked is a closed question, starting with a modal verb or an auxiliary. For example, users may wonder whether it is possible to install two versions of a product on the same machine, or whether a new product is capable of performing a specific task. The documents that contain answers to such questions are, unsurprisingly, classified as Frequently Asked Questions (FAQ) documents.

Technical support documents sometimes require both types of content - technical instructions and FAQs - especially when the initial question has more than one solution. In those cases, documents tend to point to other documents rather than including all solutions within the same document. This separation of content into reusable information chunks is a clear example of the reuse philosophy that certain companies have adopted. At Caterpillar, for instance, these chunks were called Information Elements (Hayes et al., 1996). Content developers are increasingly encouraged to use an 'author once, publish many times' approach. This approach is particularly visible for the development of user manuals and online help, whereby the emphasis is placed on sections of texts, also known as topics, rather than on actual documents. Developers are encouraged to write self-sufficient chunks of text that can be reused when needs be, regardless of the output format. The technology facilitating this approach often relies on a Content Management System (CMS) back-end, which stores these units of information.

When a knowledge base is linked to a CMS, the content then becomes available for publication within other types of documentation. This means that content written for a technical support document may be reused in a future manual or online help. This leverage opportunity offered by a CMS is optional, but shows that technical

support content may find its way into other types of documentation, which fulfils the generalisation requirement outlined earlier. From a translation perspective, the translation requirement rests on the content rather than on the final technical support document. Before dealing with the content, it is necessary to understand the sometimes complex relationships between technical support content, document, knowledge base, and CMS. Figure 1-6, outlines these relationships:
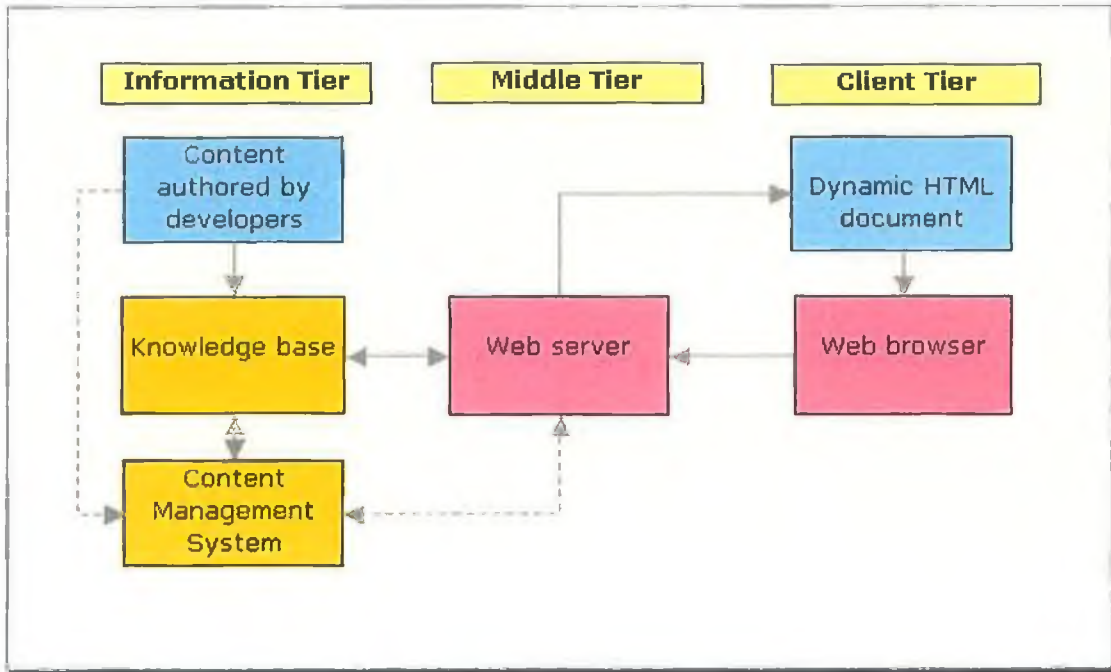
Figure 1-6: Information workflow in a Web-based architecture

This diagram indicates how technical support content may be directly authored in a knowledge base, and then published dynamically in HTML by using Cascading Style Sheets (CSS). When users request a particular technical document online, the Web server outputs the corresponding Web page based on the latest version of the content. When updates are required, developers then modify the database content, without affecting the HTML display. This explains why the same document may look slightly different from one visit to the next. Ideally, changes made to existing content are tracked within the database or by an underlying CMS, because source content updates must be translated in all target languages. In section 1.3.3, the focus is placed on the extraction of content from various knowledge bases in order to compile the corpus.

1.3.3. Compiling the corpus

1.3.3.1. Selecting popular content

As discussed in section 1.3.1, it is necessary to focus on technical support documents destined for home users, due to the assumption that translated documentation is more relevant for home users than for system administrators. However, the range of Symantec products destined for such users is quite large, with an estimated cumulative customer base of 4 million French and German users[4]. At the start of this study, approximately 30 different consumer products were available from the selection list on the technical support Web site. In order to meet the design requirements outlined earlier, the most popular products were selected to concentrate exclusively on their respective technical support knowledge bases. In terms of corpus size, the objective of 200,000 words made it necessary to target several products. This decision presented a few advantages from a representativeness point of view. It meant that the content extracted was authored by several content developers. After checking the names of the authors of various articles in the knowledge bases themselves, it turned out that the articles related to Norton AntiVirus™, Norton Internet Security™ and Norton SystemWorks™ had been authored by at least 15 different content developers. The risk of using data resulting from authors' idiosyncrasies was partly avoided. Though there may seem to be a large number of authors, the technical support content development took place over 7 years. This 7-year timeframe is also an advantage with regard to the representativeness of the corpus.

There were other reasons for selecting these three Symantec products, one reason being the popularity of Symantec's Norton AntiVirus™ product. A study conducted by the IDC, a global provider of market intelligence for the information technology and telecommunications industries, showed that Symantec held a 40.4% share of the worldwide antivirus software market (Burke, 2004: 5). In its summary report, Burke provides the following definition for antivirus software:

---

[4] Information received from Hesham Abuelata, Senior Director, Global Programs (Symantec Consumer Marketing group) on August 8th, 2006.

> Antivirus software scans hard drives, email attachments, floppy disks, Web pages, and other types of electronic traffic (e.g., instant messaging [IM] and short message service [SMS]) for any known or potential viruses, malicious code, Trojans, or spyware.

Norton AntiVirus™ is key to Symantec revenues. This is illustrated by the high number of technical support documents that have been translated into French and German: 40% and 50% respectively. These figures show that more than half the documents remain in English because of a lack of resources to translate all new source content. However, documents are only translated depending on the critical nature of the phenomenon they describe.

Two other products, Norton Internet Security™ and Norton SystemWorks™, also belong to the Norton line of products as bundled suites of applications. These suites contain applications such as Norton AntiVirus, Norton Personal Firewall, Norton Ghost, or Norton AntiSpam. Market figures are not available for suites of applications. However, Burke reveals that Norton products are pre-installed on 50 million new PCs worldwide each year (2004: 22) due to Symantec's relationship with PC manufacturers. Despite this significant figure, the translation ratio for technical support documents related to those two products is much lower than for Norton AntiVirus™. 25% of Norton Internet Security documents are being translated into French compared to 27% for German, whereas 15 % of Norton SystemWorks documents are translated into French compared to 22% for German.

For the three products, the average translation ratios are 30% for French and 35% for German, clearly indicating an opportunity to find an alternative way to translate the remainder of the documents.

1.3.3.2. Challenges of corpus compilation

The process of text extraction to build a corpus presents a number of challenges, which are described by Olohan (2004: 45). She notes that permission must be granted from the copyright holder 'to make and hold a copy of a text electronically' (2004: 50). This permission was granted by Symantec since the researcher was involved in a pilot project. The next issue she raises concerns the format in which the corpus should be stored so that it can later be processed by a particular piece of software.

1.3.3.2.1. Storage format of the corpus

The inner workings of the application chosen to extract segments from the corpus will be described in Chapter 3. This particular application required the corpus to be stored into text format. Two solutions were available to convert the HTML technical support documents in text format. The first was to access the various HTML pages via the Symantec technical support Web site, and save them as text files with the help of a Web browser. The HTML could also have been accessed and directly copied and pasted and saved as text. In these two scenarios, a tedious post-processing stage would have been required to remove unwanted tags or superfluous line feed characters that may have been inserted during the conversion process. Besides, all HTML-generated documents related to the products identified would have been difficult to locate. Due to the nature of Symantec's technical support Web site, it was not possible to get a view of all of the existing documents for a specific product. If such a list were available, it would have been possible to create a program to crawl those pages and save the resulting converted content automatically. Since an alternative method was required, special access was requested to access the knowledge base itself, similar to the access that content developers and translators have. A privileged view of all source technical content per product was obtained. For each article from the list, relevant sections were copied and pasted into a text file. A macro was then written to extract the administrative information that should not be included in the text file itself, but in a reference file. This reference file was used to track the following information:

- The internal reference number of the document
- The version names of the product
- The version names of the platform
- The first publication date of the document
- The last update of the document
- The languages in which the document is available
- The number of screenshots present in each document

All of this information was retained so as to be used as possible filters in the remainder of the study. It was also decided to add the date when the document was extracted and to mark the category in which this document was classified. When

using the privileged view per product, all documents were classified according to several categories, including installation issues, FAQs, compatibility issues, general information, functionality issues, error messages, or product-specific issues. As one document could be referenced in more than one category, it was necessary to ensure that the content was extracted only once from the knowledge base. Other challenges encountered with this approach are described below.

1.3.3.2.2. Content extraction post-processing

The requirement to save the content as text-only meant that all the database's formatting information was lost, affecting titles, bulleted lists, cell rows, and hyperlinks. From a corpus analysis point of view, this content had to be correctly identified and parsed during the analysis process. Two new line characters were required between two independent segments to ensure a correct identification of segment boundaries. For instance, a title without a final punctuation mark, such as a full stop, was separated by two new line characters to ensure that it would be analysed separately from the first sentence of the following paragraph. New line characters were automatically inserted at specific locations to separate non-related segments. These replacements were performed on all of the texts that were extracted from the knowledge base. Among the 389 documents collected, 140 concerned Norton Internet Security. 177 were related to Norton AntiVirus™, and the remaining 72 came from the Norton SystemWorks knowledge base. Specific characteristics of the corpus were identified when calculating word counts and sentence length counts.

1.3.4. Specific corpus characteristics

The size of the corpus was first calculated with a corpus linguistic tool, WordSmith Tools 4.0 (Scott, 2004), which showed that it contained 224,624 tokens for a total of 3,719 distinct types. However, the word count produced by this tool only provided a raw tokenisation of the texts. The word count was then re-run with an application whose tokeniser had been fine-tuned to deal with specific tokens such as virus names, URLs, or file extensions. This application, acrocheck™ (Acrolinx, 2005), produced a total word count of 220, 561. Table 1-1 shows examples of such tokens.

Token class	Example
URL	www.symantec.com
Directory Path	C:\temp
File extension	.gif
Virus name	W32.Sober@mm

Table 1-1: Examples of technical tokens present in the corpus

The difference between the word counts indicates that a fine-tuned application was required to handle the technical token classes that are inherent to technical support documents. This also suggests that particular attention would have to be placed on these elements when using a general purpose MT system.

The length of sentences was also calculated, and 1167 sentences were found to have more than 25 words in the corpus. Such a high number may be explained by the frequent use of long error messages and document titles embedded in the middle of sentences. For instance, a document title used in a post-modifying position is a textual characteristic of the corpus. As discussed earlier, hyperlinks establishing a reference to other documents are commonplace in technical support documents. One of the characteristics of technical support content is that these hyperlinks can sometimes be used in an appositive position, as shown in the following example:

> After you download the program, follow the instructions in the document Error: "Unable to initialise virus scanning engine. . ." after running Windows System Restore or installing Norton AntiVirus 2003/2004.

In the above example, there is no end-of-sentence character between the word 'document' and the name of the document, which is a hyperlink. Such a sentence is bound to be problematic for an MT system because its length introduces complexity and ambiguity. The absence of a punctuation mark such as a colon or a comma means that a disambiguation issue must be resolved. Even though the word 'Error' is capitalised, it may be possible to mis-parse the first part of the sentence by analysing it as: 'follow the instructions in the Error of the document:...'.

The use of appositive elements is not restricted to error messages. In technical support documentation, and in software documentation in general, User Interface (UI) options are often used as pre-modifiers, as shown by the following example taken from the corpus:

> At the Extract File dialog box, in the Restore from box, type the following where <CD-ROM drive> is the drive letter of your CD-ROM drive:

In the above example, two UI options ('Extract File', and 'Restore from') are used as pre-modifiers. When such UI options are not formatted or marked in the source text, ambiguity is introduced and comprehensibility issues may arise. This is another characteristic of software-related technical documentation that will need to be examined in detail prior to the MT process in the second part of this study.

Apart from the generation of raw numbers of tokens and types present in the corpus, no further lexical statistics were generated from the corpus. From an MT perspective, lexical acquisition is crucial to the customisation of user dictionaries, because any term that is not present in the system's dictionaries will be mistranslated or not translated at all. In order to avoid duplicating MT dictionary work, one of the first steps in designing a CL is often to standardise words and terms based on clearly-defined meanings. This approach is summed up by Nyberg and Mitamura (1996: 77):

> A key element in controlling a source language is to restrict the authoring of texts such that only a pre-defined vocabulary is utilized.

However, the aim of this study is not to define a strict controlled lexicon for a particular text type, even though it is bound to improve the quality of the MT output. In an ideal scenario, Hayes et al. (1996: 91) state that the controlled 'vocabulary database would provide the English side of a local dictionary database.' There would be merits in studying the distribution of synonyms and variants present in the corpus, and the subsequent effort involved in a systematic standardisation of such terminology. But this effort falls outside the remit of this study, since such a controlled lexicon would only pertain to Symantec documentation. In this study, it is therefore assumed that the use of unusual synonyms and term variants would degrade MT output. This assumption is based on the comprehensibility issues that

may occur when writers use new terms to refer to already-defined concepts (Warburton, 2001: 678). It may be true that new terms have to be coined to describe new ideas, especially in highly technological spheres. But what is problematic from a comprehensibility perspective is the use of synonyms, or words with unusual spelling or parts of speech. As Kirkman (1992: 97) points out, they tend 'to surprise the readers, and distract their attention', one example being 'install' used as a noun.

Instead, the focus will be placed on the extraction of segments violating specific CL rules, since these rules could possibly be used in other environments. Existing corpus linguistic tools, such as WordSmith Tools 4.0 may be used to identify basic linguistic patterns addressed by specific CL rules. One example is Bernth and Gdaniec's 21st rule, which states that the pattern *'(s)' should not be used to indicate plural'* (2001: 194). Figure 1-7 shows the results of a concordance query for the simple pattern **(s)* using WordSmith Tools 4.0's Concord module:

N	Concordance	Set	Tag	Word #	# os	# os	# os
1	scanned during a manual scan. Product(s) Norton AntiVirus 2005, Norton			527	0 527	0 527	
2	2005 3 User License Operating System(s): Windows 2000, Windows XP Date			538	0 538	0 538	
3	type. Click OK. Email Client Exclusion(s) to enter Becky! Mail This will be the			223	0 223	0 223	
4	these problems. Read only the section(s) that apply to you. To proceed if you			816	0 816	0 816	
5	alerts that you see, scan your hard drive(s) for Internet-enabled applications. For			477	0 477	0 477	
6	this message: "The following product(s) must be uninstalled before all			24	0 24	0 24	
7	Control Panel to remove the product(s). If you do not wish to remove the			52	0 52	0 52	

Figure 1-7: Query results displayed in WordSmith Tools 4.0's Concord

This approach works well for basic patterns, but is not suited to identify more complex linguistic structures on a corpus that has not been tagged. The approach used to extract violations of rules will be described in Chapter 3, once a set of CL rules has been selected for the evaluation. The rule selection process is discussed in section 2.4 of Chapter 2.

1.4. Summary

In the first part of this chapter, some of the localisation requirements for Web-based technical documentation have been reviewed. Section 1.2 showed that MT is increasingly being used to address a Web localisation bottleneck despite a lack of clear and precise Web content development rules. In order to address this issue, precise CL rules must be reviewed and evaluated empirically so that the most effective rules can be isolated. This empirical evaluation will be carried out by using genuine content destined for genuine users. In section 1.3 of this chapter, a corpus of technical support documentation has been assembled to provide the test content that will be used in the remainder of this study. Once a list of clearly-defined CL rules is compiled in Chapter 2, violations of these rules will be extracted from the corpus and used for the first phase of the evaluation process.

Chapter 2

Chapter 2: Identifying MT-oriented CL rules

2.1. Objectives of the present chapter

The concept of Controlled Language (CL) was briefly presented in the introduction. In this chapter, it is further explored to achieve two objectives. The first objective is to examine various CL evaluation frameworks in order to optimise the evaluation methodology used in this study. The second objective is to harvest a list of clearly-defined CL rules whose impact will be evaluated in Chapter 4, using segments extracted from the corpus. In order to achieve this second objective, several related concepts are discussed in section 2.2 so that the harvesting of existing CL rules focuses on the rules that are supposed to improve the machine-translatability of English source content.

2.2. Controlled Language, sublanguage, and translatability

The solution to the cross-cultural communication deadlock presented in section 1.2.1 of Chapter 1 sometimes lies in a complete automation of the translation process by using Machine Translation (MT). One of the best-known examples of an MT implementation is probably that of the domain-specific MT system derived from the TAUM-METEO project and which has been used since 1976 in order to simultaneously produce weather forecast bulletins in English and French (Isabelle 1987: 274). However, the success of this system mainly relies on the very restricted nature of the language used in the subject field of weather forecast bulletins. Such a language is defined as a sublanguage, because it uses a restricted lexicon and a limited number of syntactic structures.

A sublanguage is defined as a closed subset of sentences and words that is shared by a 'community of speakers' (Kittredge, 1982: 111). Incidentally, these patterns can sometimes hover on the verge of grammaticality, when 'they are not usual in the standard language' (Kittredge, 2003: 437). If this sublanguage is to be processed by an MT system, the MT system must be designed to handle such unusual sentences. The TAUM METEO project showed that a very specialised subject field could yield

MT output of publishable quality when using a domain-specific MT system (Lehrberger & Bourbeau, 1988: 51). But Nirenburg (1987: 12), among others, warns that it may be difficult to find a 'self-sufficient and useful sublanguage' that justifies the design of such an MT system. The main difference between a sublanguage and a controlled language is that a sublanguage is not artificially controlled, it just happens to have a limited number of linguistic features. The closed nature of a sublanguage is due to the restricted domain it describes. On the other hand, a CL is a subset of a natural language that has been specifically restricted with regard to the lexicon, syntax, and style it uses. The lexical and grammatical restrictions that define a CL are therefore the results of well-thought-out choices. These features are further described in the next section.

2.2.1. The emergence of Controlled Language

The origins of CL can be traced back as early as the 1930s with Charles K. Ogden's *Basic English* (Ogden, 1930). Ogden designed a restricted lexicon that was comprised of 850 words. This restricted language was 'intended to be used both as an international language and as a foundation for learning standard English' (Huijsen, 1998: 5). It is worth noting that Ogden's Basic English was never designed with translation in mind, but rather to solve ambiguity problems such as synonymy or polysemy for readers of English texts. Incidentally, these readers were intended to be both native speakers of English and non-native speakers of English. Ogden's ideas were then emulated a few decades later in the automotive industry with the introduction of Caterpillar Fundamental English (CFE). This CL was 'intended for use by non-English speakers, who would be able to read service manuals written in CFE after some basic training' (Nyberg et al., 2003: 261). A survey of CLs (Adriaens & Schreurs, 1992: 595) indicates that this CL was quickly followed by Smart's Plain English Program (PEP) and White's International Language for Serving and Maintenance (ILSAM). The latter gave birth to the Simplified English (SE) rule set developed by the Association Européenne des Constructeurs de Matériel Aérospatial (AECMA). The SE rule set went on to

become a standard in the authoring of aircraft maintenance documentation[5]. All of these projects had one characteristic in common: they used CLs that were designed to improve the consistency, readability and comprehensibility of their source text for human readers. These CLs are therefore often regarded as Human-Oriented CLs, and according to Lux and Dauphin (1996: 194) are not adequate for Natural Language Processing (NLP) due to their lack of formalisation and explicitness. This can be seen by some of the vagueness associated with AECMA SE's rules for descriptive writing, such as rule 6.2, *'Try to vary sentence lengths and constructions to keep the text interesting'*.

2.2.2. Controlled language and translatability

Another project worth mentioning is the collaboration between the Carnegie Group/Logica and Diebold (Hayes, 1996: 89, Moore, 2000), whose objectives were slightly different than the aforementioned projects. Instead of using CL to improve exclusively the consistency, comprehensibility, and readability of source documentation, Diebold Inc was interested in introducing a CL to also optimise its existing translation workflow, which was using translators and translation memories. The optimisation of this workflow was to be done by 'reducing wordcount, increasing leverageable sentences, and reducing the amount of expensive terminology' (Moore, 2000: 51). Despite reporting savings of 25% in translation costs with the introduction of CL, Moore also mentions that other benefits were harder to quantify, such as customer satisfaction or fewer support calls.

---

[5] AECMA SE was renamed ASD STE in 2004, after AECMA joined other European organizations to form the AeroSpace Defence Industries of Europe (ASD). SE became Simplified Technical English (STE), and the SE guide became the ASD STE-100 Specification. Since AECMA SE is the most commonly used name, it will be used in the remainder of this study.

2.2.3. Machine-oriented Controlled language

In contrast to human-oriented CLs or CLs for translation, Machine-Oriented CLs intend to improve the comprehensibility of source texts by NLP applications. (Huijsen, 1998: 2). These applications are not necessarily MT systems, as discussed in section 2.2.3.1.

2.2.3.1. Pure Machine-Oriented CL

Certain CLs have been designed to improve the comprehensibility of text by programs using logic programming or Artificial Intelligence (AI) components. Attempto Controlled English (ACE) falls into this category, being 'a computer processable subset of English for writing requirement specifications for software' (Fuchs and Schwitter, 1996: 125). By removing ambiguity from the source text, automatic validation of these specifications is possible. This strict CL is characterised by the very low number of structures permitted. For instance, only declarative sentences using third-person singular simple present verbs are allowed.

Another example is the CL that was designed at Ford (Rychtyckyj: 2006), so that 'engineers can write clear and concise assembly instructions that are unambiguous and machine-readable'. This CL, named Standard Language, restricts the number of verbs (169) and the type of sentences (imperative) that can be used by certain Ford writers. As ambiguity is removed from these task instructions, the Artificial Intelligence component of Ford's Global Study Process Allocation System can determine the 'length of time that this task will require' (ibid). The success of this CL is worth reporting because Ford is now using a general purpose MT system, Systran, to translate instructions written in Standard Language. However, Rychtyckyj reports (ibid) that certain unconventional structures allowed by Standard Language, such as 'Robot Spot-weld the Object' caused problems for Systran. This example confirms that not all ambiguity issues will be handled in the same way by NLP applications, so the effectiveness of certain CL rules may not be reproduced from one MT system to the next. This view has been expressed by Hutchins and Somers (1992: 94), who state that 'it does not really matter what kind of ambiguity the system is up against; what matters is whether the system has the relevant data for disambiguation'. When a CL is used in conjunction with an MT

system, the removal of lexical, structural, and referential ambiguity issues may be regarded as a pre-processing step in an MT workflow. Clémencin (1996: 32) states that a machine-oriented CL attempts to 'simplify and normalise the linguistic content of documents in order to match the capacities of automatic translation tools'. The next section focuses on examples of such CLs.

2.2.3.2. Controlled Language for Machine Translation

The first companies to really make use of a CL to reduce their translation costs were Rank Xerox using a SYSTRAN MT system (Adams et al., 1999: 250) and Perkins Engines (Pym, 1990) in the 1980s. Various companies soon imitated them, and one of the most successful projects to combine CL with MT was the collaboration between Caterpillar and Carnegie Mellon University throughout the 1990s. Whereas Perkins Approved Clear English (Pym, 1990) used a small number of rules (10) and a small lexicon, the CL developed at Caterpillar was characterised by its strictness. As a revamped version of CFE, Caterpillar Technical English (CTE), was specifically designed to improve the clarity of the source text so as to remove ambiguities during the automatic translation process (Kamprath et al., 1998). Despite a significant productivity hit in source authoring (Hayes et al., 1996: 86), which may be explained by interactive disambiguations that authors had to perform, the introduction of CTE's 140 CL rules and controlled terminology enabled the heavy machinery manufacturer to significantly reduce translation costs by publishing machine-translated documentation in multiple languages. The particularly high and cumbersome number of rules can be explained by the fact that the MT system used, which was based on the KANT MT system (Mitamura et al., 1991), involved an interlingua process. The abstract representation of the source text, obtained after the parsing of the English sentences, had to be universal, so as to generate sentences in multiple target languages. In order to make this process as efficient as possible, Nyberg and Mitamura (1996: 77) conclude that any ambiguity had to be resolved in the source text to ensure translation accuracy. This implementation of CL for MT showed that the accuracy of the MT output depended heavily on the level of control present in the source. The fact that such large companies committed to such a paradigm may be explained by improved communication between development groups and localisation groups. Amant (2003: 56) explains that for a long time, 'members of both fields (translation and

technical writing) perceived their professional activities as separate from one another'.

Around the same time, similar CL and MT projects at General Motors (Godden, 1998) and IBM (Bernth, 1998) showed that CL rules could significantly improve the quality of MT output in various language pairs. Other benefits were, however, more difficult to quantify. This was mentioned by Godden and Means (1996: 109), who reported that benefits such as higher customer satisfaction could not be measured but argued strongly for the implementation of the Controlled Automotive Service Language (CASL) rule set. In the present study, such a variable will not be studied directly, but findings related to the acceptability and usefulness of MT documents may help determine how satisfied users are with MT documents.

2.2.4. Development of CL rules

Determining which CL rules are the most suitable for a specific environment is, of course, complicated by the proliferation of CL rules. Many CLs and CL rules have emerged in the last two decades, whether as prototypes or as real-life implementations. In a paper presented at the fifth Controlled Language Application Workshop (CLAW), Pool (2006) reviewed CL projects carried out in the last 25 years (whether the CL was aimed at the Web, machine reasoning, machine translation, or human intelligibility), and found '41 projects that attempted to define written controlled varieties of English, Esperanto, French, German, Greek, Japanese, Mandarin, Spanish, or Swedish'. All of these CLs did not necessarily attempt to control English as source language. Examples of CLs in languages other than English include GIFAS Rationalized French (Barthe et al., 1999), ScaniaSwedish for Swedish (Almqvist and Sågvall Hein, 1996: 159), Modern Greek (Vassiliou et al., 2003: 185), or Siemens-Dokumentationdeutsch (Schachtl 1996) for German. Very often, however, the objective for the introduction of source control is similar across CL projects or prototypes: improve the comprehensibility of source documents, be it for humans or machines, by reducing or removing ambiguity and complexity from source input. On that point, one may wonder whether a machine-oriented CL is a subset of a human-oriented CL, or whether the reverse statement is true. For instance, Huijsen (1998: 2) states that certain rules, such as the one restricting the use of pronouns or ellipses, will be more helpful for machines than

for humans, and that certain rules will be more helpful for humans than for machines, such as the position of dependent clauses with regard to the main clause. On the other hand, Reuther (2003: 131) found that 'readability rules were a subset of translatability rules', which allowed her to conclude that 'translatability ensures readability, whereas the reverse statement was only true to some extent'.

This proliferation of projects suggests that CL rules vary according to the language directions, but also from one MT system to the next, or from one text type to the next. This has been confirmed by a study (O'Brien, 2003: 111), which found that eight English CLs shared only one common rule, 'the rule that promotes short sentences'. The fact that CL rule sets are so individual suggests that the introduction of existing CLs should not be performed without any evaluation. However, Hartley and Paris (2001: 322) insist that 'the corresponding middle territory between document type descriptions and sentence rules – the controlled language gap – is underexplored'. In order to be able to port CL rules from one type of document to the next, or from one environment to the next, the effectiveness of a rule must first be evaluated. Specific evaluation methods are reviewed in the next section.

2.3. Evaluation of CL rules

2.3.1. Factors hampering the evaluation of CL rules

CL rule sets such as AECMA SE focus on principles such as clarity and simplicity to remove ambiguity and complexity from source content. Yet, Nyberg et al. (2003: 276) state that 'present-day human-oriented CLs are often not specified very precisely and consistently. This causes confusion in their application and complicates evaluation.' When certain CL rules are described in generic terms, writers or content developers may apply more changes than they are required to. If developers perform undesirable and unexpected changes, it is therefore not surprising that some writing rules 'may even do more harm than good' (ibid, 2003: 105). Nyberg et al.'s statement is confirmed by some of the AECMA SE's rules (European Association of Aerospace Industries, 2001). For instance, rule 4.1 states that writers should 'keep to one topic per sentence'. Since the concept of 'topic' is not defined with objective criteria, the adherence to this rule may vary from one writer to the next. The lack of granularity in the definition of CL rules may also be explained by the insufficient linguistic background of writers and content

developers in general. If the linguistic phenomenon addressed by a CL rule is described in minute detail using linguistic terminology, content developers may not be able to implement the required change.

Their potential lack of linguistic knowledge may prevent them from understanding which word or phrase should be altered, replaced or removed. This problem has been identified by van der Eijk et al. (1996: 64) who state the following:

> Grammar restrictions often can only be expressed in a linguistic jargon that is not always easy to explain to authors, who normally are domain experts with no or limited linguistic background.

This issue is especially relevant when the CL rule is only proscriptive, because writers are not told what they are allowed to write. This issue has been raised by CL checker developers who state that some of the AECMA SE's examples 'do not always represent the best advice' (Wojcik and Holmback, 1996: 26). Uncertainties surrounding reformulations can then also arise if a CL checking application does not provide any alternative rewriting of problematic sentences. This was the case with two of the checkers that were developed to implement some of the AECMA SE's rules: the Boeing SE checker (Wojcik et al. 1990) and the EUROCASTLE checker (Clémencin, 1996: 40). Wojcik and Holmback (ibid: 23) mention that the Boeing SE Checker did not 'attempt to propose revisions of sentences', while Clémencin (1996: 40) found that the reformulation of sentences could not be fully computed based on the information generated by the linguistic analysis of the EUROCASTLE checker.

2.3.2. Evaluation of CL rule sets

Empirical studies have often focused on the effects of specific CL rule sets, such as the SE rule set developed by AECMA. For instance, Shubert et al. (1995a) focused on the effect of SE on the comprehensibility of Airplane Procedure documents using a comprehension test, while Shubert et al. (1995b) and Spyridakis et al. (1997) focused on the effect of SE on the translatability of procedural documents.

While the previous studies focused on human-oriented CL rules, Bernth (1999) focused on the impact of a set of CL rules on MT output. The usefulness of a machine-translated technical document that had been pre-edited using the EasyEnglishAnalyzer's recommendations (Bernth, 1998) was evaluated by native

speakers in the target language. She reports that the number of 'useful' translations jumped from 68% to 93% when the recommendations were taken into account prior to the MT process using the LMT MT system. A similar study was performed by Bernth & Gdaniec (2001: 208), whereby an English sample text of 69 Q&A sentences in the domain of plant care instructions was rewritten according to 13 rules. The usefulness of these principles was then evaluated at the text and sentence levels by comparing intelligibility scores given by a group of three native speakers in each target language (French, German, and Spanish). The MT output was produced by different MT systems for each language pair. Quality improvements ranging from 4% to 15% were reported at the document level, and from 25% to 36% at the sentence level (ibid).

Another study (Ramirez Polo & Haller, 2005) focused on the effectiveness of the German CL rule set used in the MULTILINT checker (Schmidt-Wigger, 1998) on the comprehensibility and post-editability of MT outputs produced by several MT systems. De Preux's study (2005) focused on the effectiveness of the CTE and SE rule sets on the comprehensibility and overall quality of MT output by counting linguistic errors. The results obtained in all of these studies were limited to specific rule sets, and did not make any claims with regard to the effectiveness of individual rules.

This overview of the evaluation of CL rule sets was limited to studies whose findings have been published. As mentioned by Bernth & Gdaniec (2001: 207), the exact results of certain evaluation studies were never published for confidentiality reasons. This is the case with the study conducted by Godden at General Motors, where the effect of the CASL rules (Godden, 1998) was evaluated on French MT output. The evaluation process was conducted by a professional translator and a bilingual automotive technician. This balanced evaluation approach seems advantageous as it involved a real user of technical documentation. In Godden's study, the impact of the CL rules on MT output was also assessed in terms of post-editing costs, and subsequent translation costs savings. This factor is often used in evaluating the effectiveness of CL rule sets, but it fails to provide a clear indication of the effectiveness of each CL rule. This gap has been filled by an empirical study

focusing on the impact of negative translatability indicators on post-editing effort (O'Brien, 2006).

2.3.3. Evaluation of individual CL rules

Due to the lack of empirical studies in the field of CL, focus is sometimes placed on theoretical linguistic issues. A shortcoming in this approach lies in the relevance of the examples chosen to demonstrate the effectiveness of some rules. It may be true that certain 'garden path' sentences are problematic for MT systems, but one may wonder how frequent and systematic those sentences are in a real-life environment. For instance, let us consider the example provided by Bernth and Gdaniec (2001: 185) to justify the application of the following rule: *'Do not omit relative pronouns; write that (which, who, etc.) explicitly'*. The example they provide without mentioning its origin is: 'The cotton shirts are made from comes from Arizona' (ibid). This example is particularly confusing, but is it representative of all the violations that this rule encompasses? This issue has been pointed out by King and Falkedal (1990: 214), who state that 'no attempt should be made to think up or import from the literature tricky examples'.

When CL rules are being evaluated as a whole, the negative impact of some may be masked by the positive effects of others. This view has been echoed by Nyberg et al. (2003: 257), who believe that 'it is unclear what the contribution of each individual writing rule is to the overall effect of the CL'. Several studies have examined the effectiveness of individual CL rules. For instance, Møller (2003) evaluated the effectiveness of 5 SE rules on the quality of MT output, but her study was not backed up empirically and lacked precise evaluation metrics. Rochford (2005) and McCarthy (2005) also geared the evaluation process towards the impact of specific rules on the quality of MT output. In her study, Rochford (2005) focused on eight rules, using between one and four examples to test the impact of each rule on the MT output of three MT systems. McCarthy (2005) evaluated more rules (22), but only used two examples to evaluate their impact on the MT quality of two target languages. These two approaches vary slightly and suggest that there is a trade-off between the number of rules that may be evaluated and the number of segments that may be used for the evaluation. When rules are evaluated individually, a test suite is often used. Such an evaluation environment is discussed in section 2.3.3.1.

2.3.3.1. Traditional test suites

Besides test corpora, another method is available to evaluate the effectiveness of CL rules on the performance or output of an NLP application such as an MT system: test suites. For instance, Baker et al. (1994) used a test suite of more than 800 sentences to assess the effectiveness of disambiguation strategies on the analyser of the KANT MT system. Their best results were obtained 'when the system was run with constrained lexicon and grammar, no noun-noun compounding and semantic restriction with a domain model' (ibid), leading to an average number of 1.5 parses per sentence, instead of 27 parses without these restrictions. Nyberg et al. (2003:273) also add that 'about 95% of the sentences were assigned a single interlingua representation.

Balkan (1994:2) defines a test suite as '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'. Lehmann et al. (1996: 711) refine this definition by stating that test suites should be systematic and exhaustive. Such an evaluation strategy appears to be particularly suited to assess the effectiveness of individual CL rules. Balkan (ibid) argues that in a test suite, 'negative data can be derived systematically from positive data by violating grammatical constraints associated with the positive data item'. This approach could be reversed to derive positive data from negative data.

2.3.3.2. Refining the test suite model

Data violating a specific CL rule could easily be turned into controlled data and both MT outputs could be compared to assess the effects of introducing the CL rule. In this case, questions may arise as to whether a single test sentence is enough to assess the effect of a specific rule that governs the use of a precise linguistic phenomenon. For instance, King and Falkedal (1990: 213) believe that the test suite should comprise 'at least two test inputs for each structure'. If an improvement is noted with the first pair of sentences, another pair of sentences using the same linguistic phenomenon should be assessed to confirm whether the improvement can be reproduced. Once a satisfactory number of test sentences has been used and evaluated for a given linguistic feature, it may be possible to generalise and conclude that the CL rule improves specific characteristics of the MT output. The

choice of specific evaluation metrics determines the manner in which MT output has been improved.

Large test suites focusing on a large number of rules and examples are probably more reliable, but they can easily become unmanageable when a human evaluation model is used. For instance, one of the CASL rules, *'Do not use more than 25 words per sentence'*, could be evaluated using an extremely large number of combinations of syntactic structures. The number of possible reformulations would be also very large since no standard reformulation exists for such a rule. Using a fully systematic and exhaustive approach is therefore not possible. A balance must, however, be achieved. This is necessary to ensure that no hasty conclusions are made with regard to the impact of a rule that has only been evaluated with a few sentences containing specific patterns. It is possible to make the impact of a CL rule more significant than it really is when specific ambiguous examples are selected. For instance, testing the impact of a rule such as *'Do not omit relative pronouns and a form of the verb "be" in a relative clause'*, with examples that only contain partitive structures, such as 'a list of addresses used', would undermine the external validity of the evaluation. The selection of CL rules used and evaluated in this study is discussed in section 2.4.

2.4. Selection of CL rules for the evaluation

2.4.1. Identification of MT-oriented CL rules or machine translatability guidelines

The main challenge with machine-oriented CL rules is that they are often proprietary and thus often unpublished. This lack of visibility reinforces the confusion that is often associated with CL rules. For instance, the CL rules that were defined at Xerox, Alcatel Bell, General Motors, Caterpillar, or Sun Microsystems have never been made completely public. Their respective authors or users (Adams et al.: 1999; Adriaens & Schreurs: 1992; Godden & Means: 1996; Kamprath et al.: 1998; Akins & Sisson: 2002) have touched on general guidelines that were used for the design of some of the rules, but without revealing the specifics of the rules. This protectionist attitude contrasts with the open approach used for the publication of general manuals of style based on in-house guidelines (Microsoft Corporation: 1998; Sun Microsystems: 2003). Certain companies

(or departments within a given company) seem to be concerned with the publication of assets they consider as valuable, regardless of the standardisation benefits they may obtain from an open policy. This trend, however, may be changing as shown by the creation of a User Group consisting of nine global companies and corporations around the concept of Controlled Language using a quality assurance application with CL checking capabilities, acrocheck™.[6]

Technical communication would benefit from a clearly defined set of CL rules, if that set of rules added value to content within a multilingual environment. In order to establish a list of CL rules to evaluate empirically, the following sources of CL rules were identified:

- Perkin's PACE (Pym, 1990)

- Alcatel's COGRAM (Adriaens and Schreurs, 1995)

- CASL rules in Bernth and Gdaniec (2001: 199)

- EasyEnglish (Bernth, 1998)

- KANT CE (Mitamura, 1999)

The type of CL rules used in the projects above vary greatly. The first one belongs to the category of loosely defined CLs (Huijsen, 1998) because its specification is not very precise. On the other hand, KANT CE's is defined as a strict CL because of its large number of well-defined rules. As mentioned previously, this set of rules is not fully available, but some examples of its principles are found in the literature (Mitamura, 1999). This statement also applies to the other rule sets presented in this list.

Besides MT-oriented CL rules, general guidelines on machine translatability are also present in the literature. Past projects (Gdaniec, 1994, Bernth & McCord, 2000, Underwood & Jongejan, 2001) have shown that there are ways to measure the machine translatability of a source text. One way to describe translatability concerns the generation of 'gross measures of sentence complexity' (Hayes et al., 1996: 90).

---

[6] For more information, see: http://www.acrolinx.com/news_release_2006_06_07_en.php

In a primitive form, this process involves counting the following phenomena and attributing some penalties:

> Sentence length, numbers of commas, prepositions, and conjunctions, supplemented by restrictions on some locally checkable grammatical phenomena, such as passive and –ing verbs (ibid)

Penalties are applied when the source text violates certain linguistic rules. The following publications in the field of machine translatability were then considered as a starting point towards the identification of specific rules:

- Logos Translatability index (Gdaniec, 1994)

- Bernth and Gdaniec's machine translatability guidelines (Bernth & Gdaniec, 2001)

- IBM's Writing Guidelines for the WebSphere Translation Server[7]

- Systran's guidelines: Preparing English Text for MT (Systran, 2005)

It should be noted that some of the rules overlap with one another. For instance, rule 22 in Bernth and Gdaniec's machine translatability guidelines (2001: 193) states that the slash character '/' should not be used to separate alternative words, as in 'user/system'. The scope of this rule is similar to the orthographical guideline provided by Mitamura (1999: 47), stating that 'the use of the slash character should be consistently specified'. Also, the rule concerning the use of explicit relative pronouns can be found in KANT CL (Mitamura 1999: 47), PACE (Pym, 1990: 85), and Bernth and Gdaniec's machine translatability guidelines (2001: 184).

Many sources, including Mitamura (1999: 47), Pym (1990: 86), and Bernth and Gdaniec (2001: 189) also recommend that writers *'avoid ellipses, except in clearly defined cases'*. Halliday and Hasan (1976: 143) define an elliptical item as something which 'leaves specific structural slots to be filled elsewhere.' According to the previous sources, relying on an MT system to fill these gaps is bound to result in incorrect parses of source inputs, which will lead to incorrect output generation.

---

[7] These guidelines were mentioned in Torrejón and Rico (2002) and available at: http://publib.boulder.ibm.com/voice/pdfs/white_papers/MT_Guidelines.pdf [Last accessed: April, 25th, 2004]. This link, however, no longer points to the document.

Since the scope of this rule is large, it was decided to split it into two rules: a rule governing the ellipsis of verbs, and a rule governing the ellipsis of subjects. It was also decided to omit lexical ellipses from the evaluation. This decision was made on the assumption that reduced forms of terms degrade MT output. For instance, if the term 'drop down' is used in source text whereas an MT dictionary contains the term 'drop down menu', the quality of MT output is bound to be affected.

The motivations that led to the design of the rules contained in the 9 sources presented earlier are not detailed, since what is particularly difficult for an MT system is well documented in the literature (Nirenburg, 1989, Hutchins and Somers, 1992, Arnold, 2003). For instance, Hutchins (2003: 505) identifies specific issues, such as 'the resolution of lexical and structural ambiguities both within languages and between languages'. Other types of ambiguity include referential ambiguities which may span several sentences. Kaplan (2003: 72) provides precise examples of syntactic ambiguity, such as dependency issues which may lead to parse failures. Kittredge (2003: 437) states that parse failures at the lexical level can be triggered by 'lexical items that have different part-of-speech or frequency of occurrence from the norm of the whole language'. After a full review and reorganisation of these rules, 54 unique rules were shortlisted. These rules are presented in Appendix A.

The review and reorganisation of the CL rules obtained from the public domain was performed using a specific strategy, whose main purpose was to isolate rules dealing with specific linguistic phenomena. None of the rules available from the public domain came with a formalised rule description, so examples in English, when available, had to be compared with one another to determine whether two rules covered the same linguistic phenomenon. In some cases, however, some of the rules collected were provided without examples and explanations, so certain rules had to be compared and classified based on their titles, regardless of the precision of these titles. For this reason, the scope of collected rules or 'negative sentence properties' (Gdaniec, 1994: 100) presented in Appendix A is sometimes much wider than the description of the final rule chosen for the present study's evaluation. For instance, the first rule evaluated in this study, *'Avoid ambiguous coordinations by repeating the head noun, or by changing the word order',* is based on Bernth's rule (1998: 33). The ambiguity issue addressed by this rule is included in the generic

problem of 'coordinations' that Gdaniec regards as a negative sentence property (ibid). Focusing on a specific phenomenon, however, was the preferred solution to try and obtain a systematic reformulation for most examples.

2.4.2. Classification of CL rules

The list in Appendix A contains rules that are designed to remove ambiguity and complexity from source content. This principle is quite generic so a more detailed classification of these rules into various linguistic categories may be required. On initial analysis, it would appear that the type of linguistic phenomenon addressed with a CL rule will impact its formal design and subsequent computational implementation. This is confirmed by O'Brien (2003: 106), who found that the classification of CL rules varies according to the linguistic framework that is used to describe a language. The classification she used in her paper divides CL rules based on lexical, syntactic, and textual phenomena. This was also the taxonomy used in the COGRAM project at Alcatel Bell (Adriaens & Schreurs, 1992). As outlined in the KANT CL project (Mitamura, 1999: 47), it is also possible to separate CL rules depending on the linguistic unit that is being addressed by the CL rule, whether it is a lexical unit, a phrasal unit, a sentential unit, or a textual unit. Both approaches imply that the classification of a CL rule depends on the level of analysis that is supported by the CL checking application used to implement CL rules. Traditionally, most checkers have operated at a sentence level. For instance, Clémencin (1996: 34) states that 'the EUROCASTLE checker works at the sentence level and has very little knowledge of the context.' This is confirmed by Bernth (2006: 1) who mentions that 'little attention has been paid to document or discourse level checking'. In a paper presented at the fifth CLAW workshop (ibid), she revealed that IBM's EasyEnglishAnalyzer was being updated to address problems occurring at the document level.

As far as this study is concerned, a CL checker will be used to extract test sentences from the corpus. This CL checker, which will be presented in section 3.2.2.1 of Chapter 3, does not perform any grammatical parse of the source text. Yet, phrase and sentence level phenomena can still be identified using a pattern matching approach. Thanks to the flexibility of this approach, it does not seem necessary to attempt to fit CL rules into specific categories initially.

2.5. Summary

In this chapter, the concept of CL was discussed from a historical perspective in order to review the various methodologies that have been used to date for the evaluation of CL rules. Several sets of CL rules and machine-translatability guidelines have also been reviewed in order to isolate the rules that have been most often implemented. This review helped us draw a list of existing rules whose impact will be analysed in Chapter 4. Before performing such an analysis, violations of these CL rules must be identified in the corpus presented in Chapter 1. The methodology used for the identification of rule violations is discussed in section 3.2 of Chapter 3.

Chapter 3

Chapter 3: Setting up an environment for the evaluation of MT segments

3.1. Objectives of the present chapter

A list of MT-oriented CL rules was identified in the previous chapter. The objective of this chapter is to set up an evaluation environment so that two sets of MT segments can be evaluated. The first set will contain violations of the rules harvested in Chapter 2, while the second set will contain segments rewritten according to specific CL rules. Section 3.2 describes the way in which a CL checker is used to extract segments containing violations of these CL rules from the corpus. The two sets of segments are then used as test data in a test suite environment. Section 3.3 discusses the parameters that were used when machine-translating the two sets of segments. Finally, the last section of this chapter describes the evaluation metrics that are employed to analyse the MT segments.

3.2. Identifying violations of CL rules

3.2.1. Data selection requirements for the test suite

As discussed in the previous chapter, designing an exhaustive test suite is not achievable within the scope of a study focusing on the evaluation of a large set of rules (54). Besides, in order to be exhaustive, one may be tempted to use artificial structures that never occur in real-life documents. For those reasons, it is preferable to follow King and Falkedal's recommendation (1990: 214), whereby 'the choice of what test inputs to include will be informed by the study of the corpus of actual texts'.

The first objective of this study is to evaluate the impact that clearly-defined CL rules have on the comprehensibility of MT segments. The test suite used to achieve this objective therefore needs to be populated with sentences from the corpus, as long as these sentences violate clearly-defined CL rules. The resulting evaluation environment should be seen as a hybrid model, borrowing from both test suites and test corpora. As discussed in section 2.3.3.2 of Chapter 2, the number of sentences extracted per CL rule will depend on the linguistic scope of each CL rule. Certain CL rules may require more test inputs than others to ensure that no hasty

conclusions are being drawn. The extraction process will also be restricted by the coverage of the corpus. Certain violations of CL rules may be absent from the corpus, but no attempt will be made to create artificial examples. The chosen hybrid evaluation environment is more flexible than a test suite as it is not based on systematicity, while the coverage of its content is broader than that of small sample texts. The approach used to extract data is described in the next section.

3.2.2. Data extraction using a CL checker

In order to identify rule violations to populate the test suite, it was decided to use a CL checker. Fouvry and Balkan (1996: 179) state that CL checkers are 'complex programs, (usually) containing parsers and (grammar) checkers, and bring in their own extra limitations and characteristics, which are a subset of the natural language'. For instance, a CL checker may flag a sentence that contains a relative clause that is not introduced by a relative pronoun. In order to find such rule violations, the checker must be pre-programmed to identify specific structures. To extract data corresponding to violations of the 54 rules selected in the previous chapter, the CL checker must be flexible enough so that simple rules can be added if they are not present by default. There is a limited number of commercial CL checkers available on the market, namely the Boeing Checker (Wojcik et al., 1990), the Maxit Checker (Smart, 2005), the Controlled Language Authoring Technology tool published by the Institute of Applied Information Science (IAI, 2002), and Acrolinx's acrocheck™ (Acrolinx, 2005). If a CL checker was to be used to extract sentences from technical documentation in the IT domain, it was essential that it could handle unusual token classes such as those identified in Table 1-1. Therefore, it was decided to use Symantec's version of acrocheck™ 2.6, which had been customised to handle such tokens. It also contained generic and Symantec-specific rules, so valuable customisation time was saved. These rules are referenced in Appendix B. acrocheck™ is presented in further detail to illustrate how it was used to extract test data for the test suite.

3.2.2.1. acrocheck's technology

acrocheck™ from Acrolinx is based on the FLAG technology described in Bredenkamp et al. (2000). Its architecture is based on a server/client model, whereby documents can be checked within a client application, such as MS Word or Adobe Framemaker via a plug-in. acrocheck™ also comes with a custom application, acrocheck™ Batch Client, that allows for several documents to be checked simultaneously. acrocheck's server contains two main parts: a linguistic software component (lingware) and a report-generating module. The report-generating module is used to store information related to the violation of rules. Once a document is checked, this information is made available to the client application. End-users are then notified of the modifications required for the document to be compliant with the predefined set of rules. The lingware component is a collection of data files used to find rule violations. Its open architecture creates an opportunity to use acrocheck™ to extract violations of rules. Specific linguistic patterns of interest can be isolated by customising rules and using acrocheck's reporting module. In order to fully understand how customised rules are created, acrocheck™'s lingware (Acrolinx linguistic engine™) must be described in greater detail.

3.2.2.1.1. Acrolinx Linguistic Engine ™

Acrolinx linguistic engine™'s technology is based on the identification of predefined patterns, which occur in a linguistically annotated string corresponding to the document that is being checked. Sentence boundaries are first identified, before sentences are tokenised to determine word boundaries. Each token present in a sentence then receives a token class depending on its linguistic attributes. For instance, a file path identified by a capital letter, followed by a colon and a back slash, and a capitalised word (such as C:\WINDOWS) receives a token class 'filepath'. Tokens with unambiguous token classes are then automatically attributed one of the 35 Part Of Speech (POS) tags from the Penn Treebank tag set, which is described in Mitchell et al. (1993) and referenced in Appendix C. In the previous example, the file path token automatically receives a proper noun (NNP) POS, because it cannot be analysed in any other way.

Once token classes have been attributed to all tokens, a tagger is used to mark each token with a POS. The tagger used is based on the statistical TnT tagger (Brants, 2000). Words missing from the lexicon can be handled thanks to a statistical approach, and ambiguous tokens from the lexicon can be disambiguated by examining contextual tokens within a three-token window. An example is shown in Figure 3-1 below:
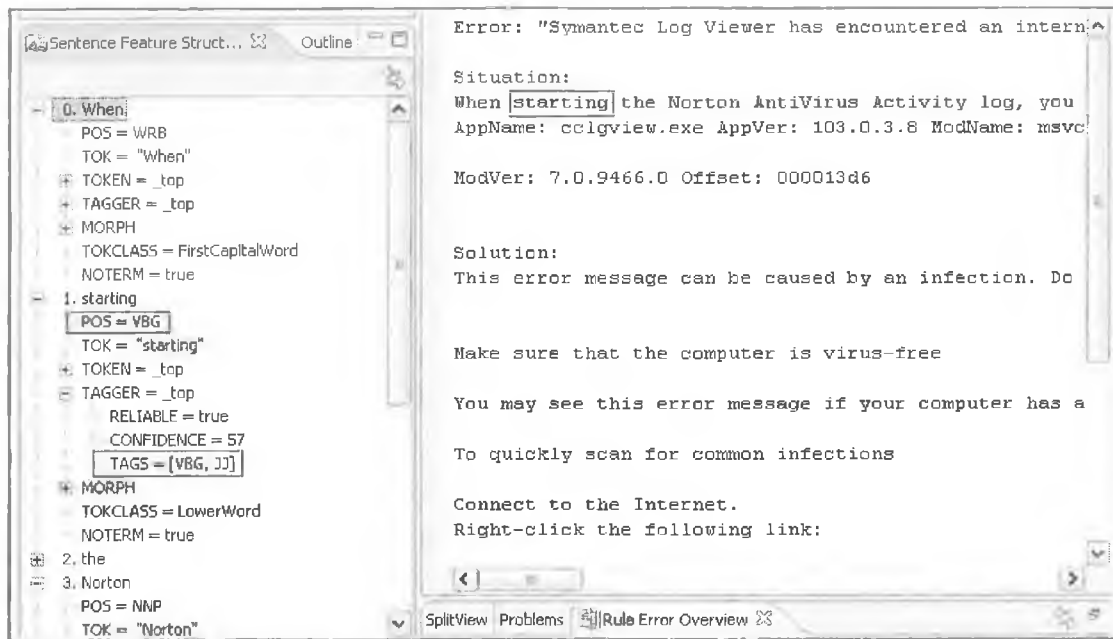
Figure 3-1: Linguistic annotation of tokens

In the above figure, the token 'starting' is an ambiguous word since it could be an adjective (JJ) or a present participle (VBG). The tagger therefore had to select one of these two POS. The upper box indicates that VBG was selected, probably due to the following word, 'the', which is a determiner (DT). A structure JJ + DT is indeed very unlikely in English, especially when the preceding word is a conjunction ('When'). Once each token has been attributed a POS by the tagger, a morphological analysis of each token is performed to provide additional information, such as the base form of a word. In short, each token of the string receives the following linguistic information:

- A token form (such as 'starting')

- A token class (such as 'Lower case word')

- Morphological information such as the lemmatised form of the token (such as 'start' for 'starting')

67

- POS with a confidence indicator (such as 'VBG')

- Term information if the term is present in a term base

Objects of varying complexity can then be created using some of these linguistic features to form patterns. The presence of illegal patterns is then checked and reported once certain conditions have been met. For instance, a comma, followed by a coordination conjunction (CC), followed by a form of 'be' could roughly indicate an instance of an ellipsis of subject. Patterns, or objects, are therefore defined first, and rules are then created by focusing on the interaction of these objects. The formalism used by acrocheck™ is summed up by Lieske et al. (2002: 6):

> The formalism can be described as a pattern matching language (including regular expressions, negation, etc.) over linguistic feature structures. A rule interpreter applies the rules to each sentence of the input text and if a pattern matches, then the corresponding rule with its assigned action is triggered.

Simple object and rule definition can be easily performed using acrocheck's Integrated Development Environment (IDE), which employs a plug-in to interface with the Eclipse Platform application (Eclipse Foundation, 2005), as shown in Figure 3-1. This environment works as a traditional acrocheck™ client application. The only difference lies in the fact that once files are checked, their full linguistic annotation data are made available so that rules can be edited and refined when 'false alarms or missed critiques' (Fouvry and Balkan, 1996: 185) are discovered. This environment provides a unique opportunity to create simple rules to complement the rules already present in Symantec's version of acrocheck™ (Appendix B). The rule creation process is described in section 3.2.2.2.

3.2.2.2. Creation of simple rules

Specific rules from the final list of 54 rules were identical in scope to some of the rules present in Symantec's version of acrocheck™, one example being rule 44, *'Avoid splitting infinitives unless the emphasis is on the adverb'*. Other rules also shared some coverage. For instance, rules 3 and 36 from the final list were encompassed by the same rule in Symantec's version of acrocheck™, *'Do not use subjectless non-finite clauses at the start of sentences'*. Dividing this rule into two separate rules was not possible due to the semantic implications of such rules. In order to extract violations of rule 3 separately from violations of rule 36, semantic information is required to determine whether the subject of a main clause is the

same as that of a non-finite clause. This limitation meant that violations of these two rules would be manually separated after the extraction process. Finally, certain rules from the final list had to be created from scratch, including the three rules governing the use of pronouns (rules 17, 18, and 19). Again, being able to separate violations of these three rules accurately would be very difficult without semantic information. Basic patterns had therefore to be created so as to extract as many potential rule violations as possible. A manual review process would then determine where they should be included in the test suite. Figure 3-2 summarises the strategy that was adopted to extract violations of such rules:
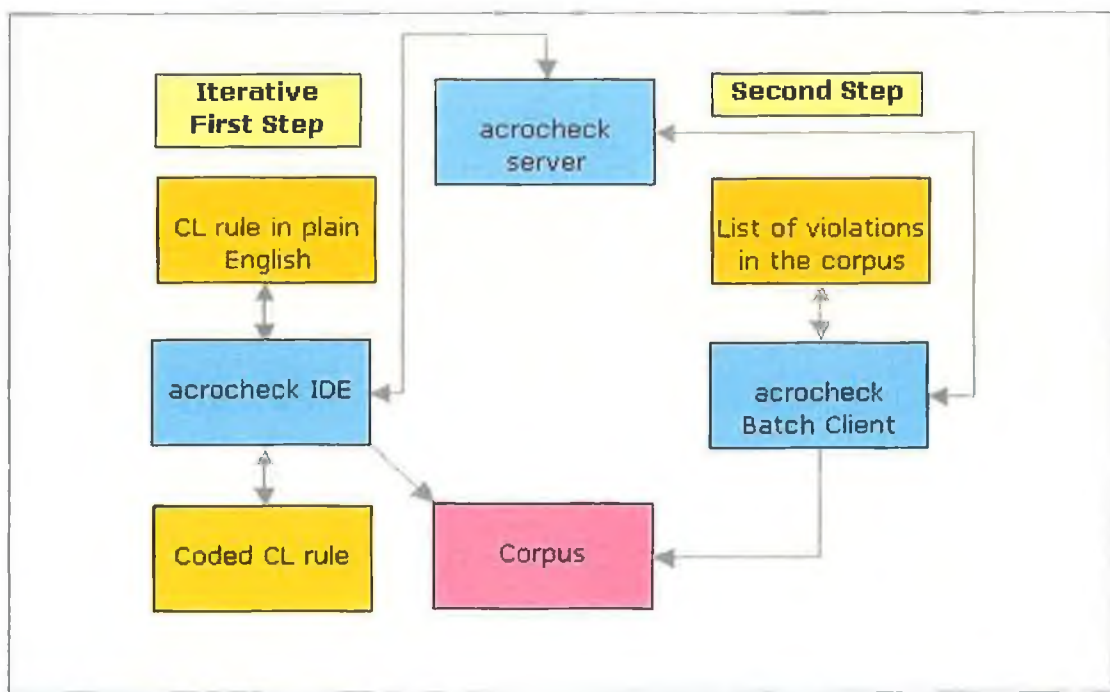
Figure 3-2: Rule creation workflow

In the above diagram, the term 'plain English' is used to refer to the descriptions of CL rules from the final list. For instance, rule 15 in plain English or layman's terms is *Do not use an ambiguous form of "have"*. When this 'rule' is coded, it will look for the auxiliary 'have', followed by an object of varying nature and length, followed by a past-participle, as in example 82: 'Symantec does not recommend having multiple firewalls installed on the same computer.' Lieske et al. (2002: 6) also found that the use of the passive voice can be easily flagged with such a formalism when any inflected form of 'be' occurred together with a past participle in one sentence. The iterative first step for the creation of such rules concerns the implementation of basic restrictions to ensure that genuine patterns are returned. These rules were

tested separately by checking several texts from the corpus using acrocheck™ IDE. Once the precision of the rules was improved to avoid too many false positives, they were used as unique rules on the whole corpus using acrocheck™ Batch Client. Issues concerning precision and recall are discussed in the next section.

3.2.2.2.1. Precision and recall

Due to the nature of the task, it was not always possible to create rules with high precision, whereby only genuine violations of the rules would be flagged. The accuracy of CL checkers is often measured with precision and recall scores. Precision is measured by comparing the number of correctly flagged errors against the total number of errors flagged. On the other hand, recall is measured by comparing the number of correctly flagged errors with the total number of errors actually occurring (Adriaens & Macken, 1995: 126). Trying to improve one score has consequences on the other. The main objective of this task was to identify violations of CL rules that contained different linguistic elements to avoid populating the test suite with similar patterns. Low precision was therefore not an issue since a manual review process would be performed to remove false positives. It is worth insisting that such 'rules' could never be used in a real authoring environment before being fully tested and refined. The consequences of implementing CL rules that have not been fully tested with data on which they are used must not be underestimated. Wojcik and Holmback (1996: 30) found that technical writers using the Boeing SE checker were more frustrated by errors in precision than errors in recall. This was confirmed by Barthe (1996: 50) who insists on the importance of updating rules to avoid 'unjustified errors', be it at the lexical or syntactic levels.

3.2.2.2.2. Challenges in rule creation

The second challenge to be addressed with regard to rule design concerned certain default settings of acrocheck's linguistic resources. For instance, rule 48, *'Question marks should only be used in direct questions'*, proved difficult to implement without changing acrocheck™'s sentence splitting rules. If a question mark is used in the middle of a sentence, an MT system may segment the sentence incorrectly. This was noted when translating the following sentence with Systran 5.0:

☒ In the drop-down menu, under the "What do you want to do?" pane, choose Manually, and click OK.

🅰 Dans le menu déroulant, sous le « Que souhaitez-vous faire ? » le volet, choisissent manuellement, et cliquent sur OK.

In the above example, the question mark is used in a direct question, but it does not stand for an end of sentence character, because the whole question qualifies the word 'pane'. In order to identify similar instances of this phenomenon in the corpus, a rule was created using acrocheck™ IDE. However, a linguistic phenomenon that is problematic for one NLP application is bound to create problems for other NLP applications. This issue was highlighted when attempting to code such a rule by examining patterns which consisted of noun objects following a question mark character and optional quotation marks. When checking the segmentation of the above example, it appeared that the question mark character had been interpreted as an end of sentence marker. By default, sentence boundary markers were inserted automatically after question mark characters. This meant that acrocheck's sentence splitting rules had to be edited so that the CL rule could be implemented. Once this issue was resolved, it was possible to use the rule on the entire corpus to extract violations of this rule. The review process of these violations is described in section 3.2.2.3.

3.2.2.3. Populating the test suite with raw segments

acrocheck™ was used to extract test sentences (or groups of sentences) from the corpus. These extracted sentences are referred to as raw segments, which were manually reviewed to obtain a final list of segments to include in the test suite. Following King and Falkedal's recommendation (1990: 213), at least two segments had to be selected for each of the 54 rules[8]. For certain rules, it was necessary to include more than two raw segments because of the productive nature of the rule. For instance, the scope of rule 12, *'Use the active voice when you know who or what did the action'*, is well-defined and its reformulations are obvious. As discussed in section 3.2.2.2, a form of 'be' followed by a past-participle and 'by' will return most violations of this rule. Reformulating such a sentence can be done consistently in

---

[8] All of the examples used in the test suite are provided in Appendices F and G.

most cases, since the agent must be turned into a subject. However, the length of the agent can vary greatly, being sometimes composed of a single word (as in example 64, 'Windows') or of a long noun phrase (as in example 67, 'a specific registry entry that was not deleted during the uninstallation of a previous Norton program'). In order to evaluate the impact of this CL rule effectively using different structures, it was decided to extract as many types of agents as possible while keeping in mind that the test suite's content would have to be manually evaluated and analysed. Twelve raw segments out of the 244 violations of this rule found in the whole corpus were therefore included in the test suite. It must be stressed that these raw segments were not randomly included in the test suite for two reasons. First, some of the 244 violations found in the corpus could have been false positives. Including random segments in the test suite would have affected the validity of the experiment. Second, the issue of structure diversity mentioned in this section would have been compromised. Certain rules therefore contain three or four times the number of examples of others in the final test suite. In order to ensure that all CL rules receive the same treatment during the evaluation, an average score will be calculated based on the number of segments they were assessed with. This will be further detailed in section 4.2.2.1 of Chapter 4.

Another criterion used for the selection of raw segments was that they should be as context-independent as possible in order to facilitate the machine translation process and the subsequent evaluation by human evaluators. Segments are often parsed at a sentential level by MT systems, but preliminary tests revealed that MT output could differ slightly depending on the context. For this reason, contextual sentences were also included in raw segments in which anaphora resolution issues were apparent. For instance, if the first subject or object of a sentence was expressed by an indefinite third-person pronoun ('it' in example 98), the previous sentence including the referent was added to the raw segment to provide the MT system with an opportunity to resort to a disambiguation strategy.

Finally, one last step had to be performed before raw segments could be reformulated into post-CL segments. King and Falkedal (1990: 213) state that 'test inputs designed primarily to illuminate the system's treatment of specific source language phenomena should avoid the introduction of "noise" triggered by an

injudicious choice of lexical material'. This recommendation was also applied to superfluous syntactic structures that have no direct impact on the problematic linguistic feature addressed by the CL rule. In the present study, the effectiveness of the CL rules will be evaluated by human evaluators. Test inputs, and corresponding MT outputs, should therefore be rid of most elements that could make the evaluation process more cumbersome as long as they have no direct influence on the linguistic phenomenon under study. The following example shows how this strategy was used:

- Rule 8: *'Do not omit the relative pronoun and a form of the verb 'be' in a relative clause when the post-modifier is an -ing word'.*

- Raw segment: If you still have the problem, then it can be caused by an open driver or by an application not **responding** to the shutdown request.

- Elements with no direct influence on the linguistic phenomenon addressed by the rule: conditional clause, pronoun

- Pre-CL segment: The problem can be caused by an open driver or by an application not **responding** to the shutdown request.

- Post-CL segment: The problem can be caused by an open driver or by an application *that is* **not responding** to the shutdown request.

Other examples of modifications performed on the raw segments to obtain pre-CL segments are presented in Appendix D. The semantic scope of certain segments was sometimes reduced or expanded depending on the type of modification performed. As far as the objective of this study was concerned, a trade-off was required between having genuine test sentences and simpler source text sentences so that the evaluation process would not be hampered by unnecessarily complex sentences. It should be noted that this step was limited to well-defined replacements or removals. As each CL rule had to be evaluated separately, modifications based on other clearly-defined CL rules were not introduced. For instance, one could have decided to ensure that all of the passive forms present in raw segments were turned into active forms in pre-CL segments. Avoiding passive forms is indeed a CL rule that is often found in the literature (Bernth & Gdaniec, 2001: 190, Farrington, 1996: 18). But what if this CL rule sometimes does more harm than good? In such a case,

some of the test suite's segments could be contaminated and the internal validity of the study would be undermined. Focusing on one change at a time is similar to the approach taken by McCarthy (2005: 12), who evaluated the effectiveness of eight CL rules by using sentences 'containing only one problem which the CL rule would aim to address'.

Once the raw segments were rid of 'noise', the resulting pre-CL segments were manually turned into post-CL segments by following the rewriting instructions of the CL rules when they were available in the final list of rules. Most of the uncertainty surrounding CL rules stems from the vagueness associated with the reformulations that writers or content developers are expected to make. A key concern in the present study was therefore to ensure that rewritings were as consistent as possible so that the impact of the rule did not vary too much from one reformulation to the next. For this reason, pre-CL segments were all turned into post-CL segments by the researcher. It would, of course, have been interesting to study whether different people would have used the same reformulation for a given linguistic feature. However, it was felt that such a study was beyond the scope of the present research. Besides, using consistent reformulations was sometimes impossible due the scope of the rules. For instance, rule 2, *'Do not use ambiguous attachments of non-finite clauses (present-participles)'* encompasses several ambiguous uses of –ing words introducing non-finite clauses. Reformulations therefore differed in examples 10 and 12 selected to test the effectiveness of this rule:

☒ Fixes computer problems using Norton AntiVirus.

☒ You need more assistance removing a threat.

In these two examples, the nature of the clause introduced by the –ing word is different. In example 10, the –ing word introduces a clause of manner that requires the preposition 'by' to remove the ambiguity. On the other hand, example 12 contains a non-finite purposive clause that requires an infinitive verb introduced by 'to' to remove the ambiguity. Once these modifications were performed, the sets of 304 pre-CL and post-CL segments were ready to be machine translated. Section 3.3 of this chapter describes the parameters that were used for this process.

3.3. Using an MT system to translate Pre-CL and post-CL segments

As outlined in the introduction, the main objective of this study was to focus on the French and German output produced by a single MT system, Systran Web Server 5.0, rather than to compare the effectiveness of CL rules on the output of several MT systems. The way in which this system was used is described in section 3.3.1.

3.3.1. Determining MT parameters

According to Arnold et al. (1994: 163), an important factor in the evaluation of the linguistic quality of MT output is the context in which the system is evaluated, either in a black-box context or a glass-box content. This parameter is discussed in the next section.

3.3.1.1. Using an MT system in a Black box context

White (2003: 225) states that a 'glass-box view looks inside the translation engine to see if its components each did what was expected of them in the course of the translation process.' On the other hand, a black-box context applies when assessment can only be performed by working with inputs and outputs (Arnold et al., ibid). An example of glass-box evaluation is found in Baker et al. (1994: 90), who 'experimented with the KANT analyzer in order to determine the effects of different disambiguation strategies'. In such a setting, the MT system generates logs of all operations performed, including the disambiguation strategies used during the analysis of source input. In terms of CL rule effectiveness, a glass-box approach is advantageous because it may allow researchers to check whether a specific CL rule has removed ambiguity from source input, and avoided complex disambiguation strategies. Within a black-box context, examining MT output cannot provide this level of information. It may be possible to determine whether the source input was correctly analysed, but it is impossible to know the confidence with which the MT output was generated. The system used in this study, Systran Web Server 5.0, was used in such a context. This MT system, however, is shipped with a wide variety of client applications that may be used to request translation jobs from the MT engine. Since the test suite assembled for the first part of this study was stored in an MS Excel spreadsheet, it was decided to translate its contents with the Systran plug-in for MS Excel. This meant that a single environment was used for the whole

evaluation process. Once the segments were machine-translated, an extra column was added to the spreadsheet so that scores could be attributed to every segment by the evaluators. The translation process was performed by making sure that Systran's translation options were set to reflect the style of the content to be machine-translated. For both language pairs, the polite form was used with the second person pronoun set to 'plural', as shown in Figure 3-3.
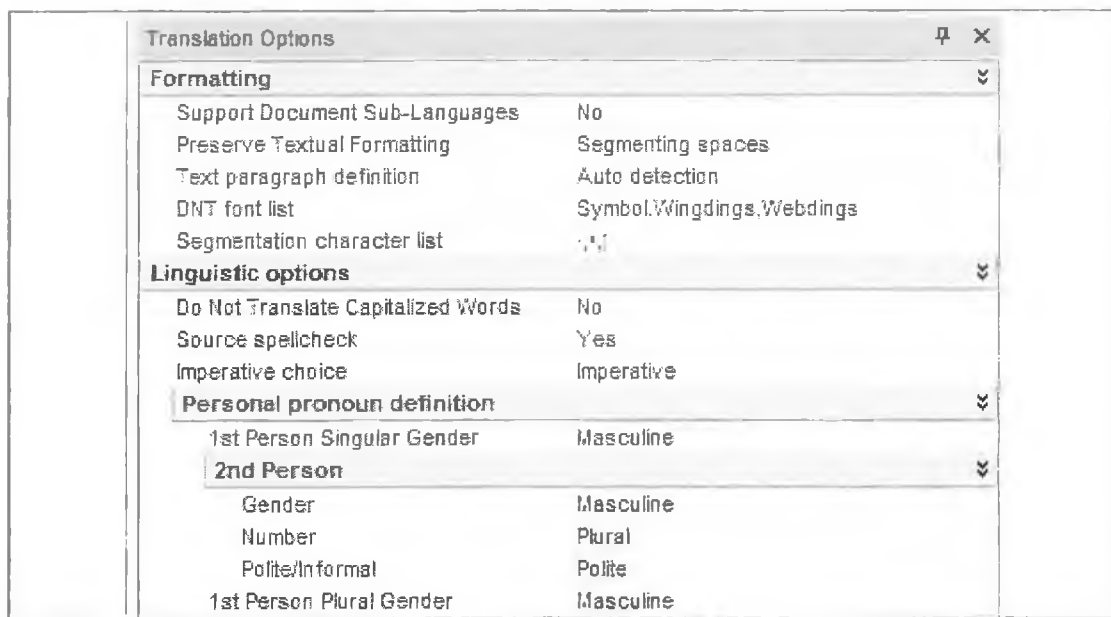
Figure 3-3: Selection of Systran's Translation Options

For both language pairs, it was also decided to use the imperative mode by default, since technical support content is often characterised by procedural steps that users are invited to follow. However, one option was specific to the English to German language pair: the activation of the new German spelling.

Lastly, specialised dictionaries were selected to ensure that MT output would not be meaningless. To achieve that objective, Systran's Sciences dictionary was activated using the Computers/Data processing domain. Symantec's own user dictionaries were also used once they were customised to handle any terminology specific to the test sentences chosen for the evaluation. The advantages and drawbacks of this approach, as well as the issues encountered during the customisation process, are discussed in section 3.3.1.2.

3.3.1.2. Using refined user dictionaries

Before the contents of the test suite could be translated and evaluated, domain-specific terminology had to be identified and added to the system's User Dictionaries. This section describes the approach taken to refine Symantec's User Dictionaries before the final translation process was conducted for the two sets of segments contained in the test suite. In the remainder of this dissertation, the set of translated pre-CL segments will be referred to as MT output A and the set of post-CL segments as MT output B. Both sets of segments were machine translated a first time to identify mistranslations of technical terms, or absence of translations for unknown terms in both outputs. Certain individual terms and compounds were incorrectly generated by the MT system for both the set of pre-CL segments, and the set of post-CL segments. It was therefore necessary to add them to the user dictionaries before performing the final evaluation. To perform this task, Nyberg and Mitamura's recommendations (1996:79) were followed:

> Whenever possible, a system should parse longer strings of technical words as single, atomic units of meaning rather than compositionally-derived structures. (...) Phrasal verb-particle constructions such as 'abide by' are also easier to analyze if taken as a unit.

This last point is exemplified by one of COGRAM's pair of approved lexical terms, 'set up', and disapproved term, 'establish' (Adriaens, 1996: 226). If the phrasal verb is to be preferred to the single form of a verb, then it must be present in the MT system's dictionary. To ensure the internal validity of the experiment, it was essential that both outputs received the same treatment, and that unknown terms occurring in both outputs would be added to the dictionaries. Bryman and Cramer (2001: 9) point out that 'an internally valid study is one which provides firm evidence of cause and effect'. If a causal relationship were to be established between the improvement of the MT output and the application of a CL rule, both MT outputs should be handled in the same way. A list of unknown or mistranslated terms was compiled and translated into both French and German. This list of terms was then imported into Symantec's user dictionaries and coded using Systran's Intuitive Coding technology described in Senellart et al. (2001: 5). The main advantage of this approach was the speed at which terms were imported and coded, where other MT systems may have required additional linguistic information. Certain entries generated problems during the coding process, but they were easily

identified by Symantec's Systran user dictionary experts who added special intuitive or coding clues (Systran, 2005: 146) to ensure that the system would use the entry properly. The following example demonstrates the use of coding clues for an English to German entry:

☒ point to > zeigen auf

☑ to point to > zeigen auf (governs_accusative)

In the example marked with a tick, two clues have been used. First of all, the source token 'point' is recognised as a verb during the coding process thanks to the word 'to' which precedes it. Besides, the German preposition 'auf' of this particular verb needs to be followed by an accusative form, hence the presence of the second coding clue within brackets (governs_accusative). The second clue is an example of an advanced language-specific coding clue[9].

This approach showed certain limitations, however, as some entries were still used incorrectly in the target text for no apparent reason. It would have been possible to refine the user dictionary entries over and over again, to try and achieve 100% accuracy, but this process would have been too time-consuming. Besides, in certain cases, it is not possible to obtain the desired result, especially when the form of the target word does not match that of the source word. This was the case with the verb pair 'to click' and 'klicken auf', which was not used as intended in the German output. By default, the German preposition 'auf' was not generated by the MT system when the English verb 'click' was used without the corresponding preposition 'on'[10]. It was therefore decided to hardcode the preposition 'auf' in the user dictionary's entry. This preposition, however, should be located in various places in the target text depending on the type of sentence, be it imperative or purposive. As it was hardcoded in the user dictionary entry, it always followed the verb, which impacted on German word order in many cases. After spending some

---

[9] The Coding Reference Table for German is available at:
http://www.systransoft.com/support/Dicts/Tables/latest/de.html [Last accessed: 25/06/2006]. When last accessed, this table did not include the following clue: (governs_accusative).

[10] This issue was reported to Systran and was fixed in a subsequent version of Systran Web Server delivered to Symantec, with the implementation of a specific coding clue.

time trying to find a solution to tweak the entry, it was decided that it would be easier to perform an automatic replacement before submitting the output to the evaluators. The underlying concept behind this replacement is described below.

3.3.1.2.1. Automatic modifications

As mentioned previously, recurring problems can impact on the MT output in an unexpected manner; an example being the verb pair 'to click', 'klicken auf'. As mentioned by Elming (2006: 219), 'one of the great advantages of rule-based MT as opposed to statistical MT is its transparency, (...) which makes the errors very consistent'. These errors may be reported to the MT system developer, but it may take some time for a solution to be found. In the meantime, two solutions are available to the MT end-user: accept the error as a PE task and rectify it during that process, or find a way to automatically correct the problem. The second approach ranges from pattern matching replacements (Senez, 1998: 294) to transformation-based corrections (Elming, ibid). A simple solution was adopted to ensure that a common verb, such as 'to click', would not affect the overall results of the evaluation. The test suite contained thirty occurrences of the word 'click', so thirty examples were potentially affected by this issue. Evaluation scores may have been affected because of this grammatical inaccuracy. Before the MT output was submitted to evaluators, a global replacement was created and performed on the German output, by restoring the pattern (\b([Dd]oppel-)?[Kk]licken\b) (\bauf\b) (\bSie\b) as $1 $3 $2. Once this replacement was performed using a macro, the MT output was ready to be evaluated.

3.4. Evaluation of the MT segments

Almost forty years after the seminal report of the Automatic Language Processing Advisory Committee (ALPAC: 1966) was published, White (2003: 212) points out that 'reliable, efficient and reusable methods' to evaluate MT output quality are still difficult to find. This is due to the fact that 'assessing translation quality is not just a problem for MT: it is a practical problem that human translators face (...) due to the number of many possible translations' (Arnold et al., 1994: 161). In order to bypass the issues that are inherent to human evaluations, several automatic evaluation methods have been developed in the last number of years. Most of these automatic evaluation methods focus on the similarity or divergence existing between an MT

output and one or several reference translations. In order to select the best-suited approach for the present study, the advantages and drawbacks of both automatic and human evaluation strategies are discussed in the next section.

3.4.1. Review of existing evaluation metrics

Whether an evaluation of MT output uses human or automatic metrics, choices must be made with regard to the selection of these metrics as no 'evaluation requirements or standards for MT output exist' (Gough, 2003: 113). The selection of specific metrics depends on various factors, ranging from time and budget constraints to the nature of the study's objectives. Prospective metrics for an automatic evaluation and a human or manual evaluation are reviewed in sections 3.4.1.1 and 3.4.1.2 respectively.

3.4.1.1. Review of automatic evaluation metrics

Examples of automatic metrics include the Word Error Rate (WER) metric described in Jufrasky & Martin (2002: 271) and Dabbadie et al. (2002: 10); BLEU (Papineni et al., 2002); NIST (Doddington, 2002); and Precision and Recall (Turian et al., 2003). Using automatic evaluation metrics that employ reference translations would require at least one set of reference translations produced from both pre-CL and post-CL segments, and comparing the results obtained for both sets. The first stumbling block of such an approach would concern reference translations. Should post-edited versions of refined MT output be used as reference translations? Or should human translations be produced from scratch? Studies using automatic evaluation metrics tend to focus on reference translations produced by human translators. For instance, Coughlin (2003: 65) used Microsoft Translation Memories (TM), and Gough (2003: 114) used TMs provided by Sun Microsystems. A slightly different approach is described in Hǎjic et al. (2003: 163), whereby reference translations are produced by post-editors using refined MT output contained in a TM.

Regardless of the approach taken, automatic evaluation metrics are often said to be an inexpensive alternative to human evaluation (Papineni et al., 2002: 318). Previously, we saw that new sets of data require reference translations, which might be more expensive to produce than performing a manual evaluation of the MT output. Popescu-Belis et al. (2002: 17) mention that this budgetary issue can become more problematic if several reference translations are used. Automatic metrics such as BLEU and Precision and Recall support multiple reference translations per source string, so several translations could theoretically be employed to evaluate the quality of the MT output. Such an approach, however, would make the evaluation costs spiral without guarantee of any benefit. Gough (2003: 113) states that 'the number of reference translations can affect the scores obtained in an automatic evaluation', which shows that automatic metrics are not entirely reliable. This view has been echoed in a study performed by Turian et al. (2003: 8), whose 'most important finding is that, even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are still very far from being able to replace human judgment.' This statement contradicts the conclusions expressed in Coughlin (2003: 69) and Papineni et al. (2002: 318). Papineni et al. state that 'BLEU's strength is that it correlated highly with human judgments', and Coughlin describes it as a 'highly reliable alternative to human evaluation'. These statements cannot be confirmed in all evaluation situations, as it transpires that BLEU's scores are more useful for large corpora rather than single sentences (Papineni et al., 2002: 318). The purpose of this study is to focus on a spectrum of improvements provided by CL rules, and human judgements appear to be a more reliable solution. The next section reviews the metrics traditionally used within human evaluation frameworks, based on the objective of the present study.

3.4.1.2. Review of human evaluation metrics

The human evaluation of MT output quality is often regarded as a perilous exercise due to the subjectivity and bias that is inherent to each evaluator. White et al. (1994: 193) stress that 'judgments as to the correctness of a translation are highly subjective, even among expert human translations and translators'. This problem is aggravated by the fact that certain evaluators may be biased when evaluating MT output. For instance, if these evaluators are professional translators, they may regard

MT as a threat. Regardless of the evaluation metrics used, this factor will consciously or unconsciously influence the way they score MT output. It is therefore necessary to use a group of evaluators to maximise the reliability of the results, and decide on well-defined evaluation tasks that these evaluators should perform.

Van Slype (1979:12) states that 'translation quality is not an absolute concept, and that it has to be assessed relatively, applying several distinct criteria illuminating each special aspect of the quality of the translation'. A wide range of criteria have been used to date, and finding those that correspond best to a given study largely depends on the objectives of this study. Van Slype (ibid) mentions that these criteria can operate at several levels:

- at the cognitive level, when the purpose of the study is to evaluate the intelligibility, the accuracy, or the usability of the MT output

- at the economic level, when the objective is to measure the effort, or the time that would be required to bring the MT output to a certain quality level.

In the present study, the objective is to isolate CL rules that will prove efficient on both counts. Ideally, they will allow for the output to be sufficiently accurate and comprehensible for users, so that the post-editing process can be bypassed. This dual objective must therefore be reflected in the criteria that will be used by the group of evaluators scoring the MT output.

3.4.1.2.1. Counting linguistic errors

One possible evaluation method is to count and compare the number of translation errors in both MT outputs. Counting errors in MT output is an approach suggested by Weissenborn in Van Slype (1979:102), Arnold et al. (1994: 164), and Flanagan (1994). It is also favoured by methodologies such as the Blackjack method, (ITR Ltd, 2002) or the Society of Automotive Engineering (SAE)'s J2450 metrics (2001). This approach has been used in studies focusing on the evaluation of the effectiveness of CL rules, by De Preux (2005:3) and McCarthy (2005: 26). However, this approach presents several disadvantages. Firstly, counting the number of errors is not totally objective since it introduces judgment calls. For

instance, the SAE J2450 metric classifies errors as either 'serious' or 'minor', but acknowledges that it is 'impossible to define the notion of a serious or minor error' (SAE, 2001: 2). This remark goes against their stated objective: 'to establish a consistent standard against which the quality of translation of automotive service information can be objectively measured' (SAE, 2001: 1). Secondly, it may be possible to check whether the number of errors is lower in MT output B, and conclude that the application of the CL rule was effective. However, it would not tell us how effective it was with regard to the comprehensibility of the output. Finally, one should not underestimate the amount of time required for evaluators to become familiar with these error categories (SAE J2450 only use 7 categories, but Flanagan and Blackjack use more than 20). Consistency in counting and classifying errors may be difficult to achieve even with clear instructions. This challenge was reported by McCarthy (2005: 30), whose evaluation study was hampered by the difficulties encountered by evaluators in categorising errors due to the 'subjective nature of the task as well as the "knock-on" effect of the different errors'. A more practical and efficient approach must be found to avoid the steep learning curve associated with such an evaluation strategy.

3.4.1.2.2. Intelligibility and fidelity

Two of the most frequently used human evaluation metrics of MT output quality are intelligibility and fidelity. For T.C. Halliday, comprehensibility and intelligibility are synonymous terms and refer to the 'ease with which a translation can be understood, its clarity to the reader.' (quoted in Van Slype, 1979: 62). Distinctions are sometimes made between the two terms, depending on the type of material that is being evaluated (isolated segments or full translation of a document). Arnold et al. (1994: 161) hold that the 'intelligibility of a translated sentence is affected by grammatical errors, mistranslations and untranslated words.' This has been confirmed by experiments conducted by Reeder (2004: 231). She discovered that some error types are immediately apparent and predicators of intelligibility. These include 'incorrect pronoun translation, inconsistent preposition translation and incorrect punctuation.' The introduction of CL rules in the source text should reduce these comprehensibility issues in the target text. Besides, a machine-translated sentence can sometimes be very difficult to understand when it contains ambiguity. For this reason, disambiguation is often regarded as a common

translation task to be performed on the target text (Loffler-Laurian, 1996: 86). As certain CL rules remove ambiguity from the source text, ambiguity should also disappear from the target text.

On the other hand, fidelity is concerned with the accuracy of the information conveyed (White, 2003: 216). While being grammatical and comprehensible, a sentence may be inaccurate if ambiguous words have been mistranslated. These two metrics fall into the category of intuitive judgments mentioned by White et al. (1994: 193): 'Evaluation must exploit intuitive judgments while constraining subjectivity in ways that minimise idiosyncratic sources of variance in the measurement.' However, intelligibility and fidelity are often measured with graded scales, which may impact the objectivity of the results. The other drawback of using these two metrics separately is that it extends the time required for the evaluation process, which in turn, increases the cost of the operation. When using two separate metrics, Coughlin (2003: 64) mentions that one has 'to determine the importance of one characteristic over another when deciding what acceptable quality is'. In her study, which focused on the correlation between automated and human assessment of Machine Translation quality, she asked evaluators to use a unique scale of 4 values to measure the acceptability of the output. This simple approach integrated criteria concerning both intelligibility and accuracy characteristics, but was easier to use and process than if the two criteria had been evaluated separately.

Finding appropriate metrics for MT evaluation based on specific requirements is a process on which the Framework for Machine Translation Evaluation (FEMTI) has been working since 2002 (Hovy et al., 2002). FEMTI is an initiative of the International Standards in Language Engineering (ISLE), whose aim is to classify various methodologies used for the evaluation of a wide range of MT systems' characteristics, including the quality of MT output. Popescu-Belis et al. (2005: 6) state that the 'most original aspect of FEMTI is a mapping from the context of use of the MT system to the quality characteristics' that must be measured with well-defined metrics. Since MT output is used for dissemination purposes in this study, FEMTI 2005 suggests that the most important evaluation metrics are accuracy and well-formedness. If the MT output is ungrammatical or contains punctuation errors, it will be more difficult to understand. Likewise, if the translated information is not accurate, users may do more harm than good to their computing environment when

trying to fix a problem using the machine-translated document as reference material. Evaluating MT output using metrics incorporating references to accuracy and well-formedness therefore seemed essential. Even though FEMTI also mentions that style is more important than speed when MT is used for dissemination purposes, the strict translation turnaround requirements outlined in section 1.2.1 of Chapter 1 prevented style from being taken into consideration. It was decided to use a similar approach in the present study based on the type of content contained in the corpus, and the context of the use of MT. Since this technical content is perishable and likely to be updated on a regular basis, translations must be obtained quickly so as to be disseminated to a large pool of users. The approach combining accuracy, comprehensibility, and well-formedness metrics, and the scale used are described in further detail in section 3.4.2.

3.4.2. Selection and design of metrics

The main drawback of Coughlin's approach is that her metrics attempted to measure the acceptability of a set of translations produced by various MT systems, without taking users into account. Van Slype (1979:13) stresses that 'acceptability can be effectively measured only by a survey of final users.' It is interesting to note that a sentence can be 'comprehensible, given enough context or time to work it out,' (Coughlin, 2003: 64) by an evaluator who has had access to the source text, but what about end-users who rely exclusively on the translated material? The metrics chosen in the first part of the present study included this requirement, based on the decision to involve genuine users in the second part of the study. The scale containing four separate values is presented in Table 3-1:

Score	Criteria
Excellent MT output (E)	Read the MT output first. Then read the Source Text (ST). Your understanding of the MT output is not improved by the reading of the ST because the MT output is satisfactory and would not need to be modified: it is syntactically correct and it uses proper terminology, even though it may not be stylistically perfect. However, the MT output performs its primary function, which is to convey information accurately. An end-user who does not have access to the ST would be able to understand the MT output.

Score	Criteria
Good MT output (G)	Read the MT output first. Then read the source text. Your understanding of the MT output is not improved by the reading of the ST even though the MT output contains minor grammatical mistakes (word order/punctuation errors/word formation/morphology). You would not need to refer to the ST to correct these mistakes. An end-user who does not have access to the source text could possibly understand the MT output.
Medium MT output (M)	Read the MT output first. Then read the source text. Your understanding of the MT output is improved by the reading of the ST, due to significant errors in the MT output (textual coherence/ textual pragmatics/ word formation/ morphology). You would have to re-read the ST a few times to correct these errors in the MT output. An end-user who does not have access to the source text could only get the gist of the MT output.
Poor MT output (P)	Read the MT output first. Then read the source text. Your understanding only derives from the reading of the ST, as you could not understand the MT output. It contained serious errors in any of the categories listed above, including wrong POS. You could only produce a translation by dismissing most of the MT output and/or re-translating from scratch. An end-user who does not have access to the source text would not be able to understand the MT output at all.

Table 3-1: Scores used by the evaluators

As shown in Table 3-1, the scope of these metrics is not restricted to a specific quality criterion. This will allow for an assessment of the MT output on several levels, combining criteria that others have decided to separate. For instance, Ramirez Polo and Haller (2005: 7) decided to separate comprehensibility and post-editability within their comparison evaluation of three MT systems. The scores chosen in the present study, however, will provide indications on the comprehensibility of MT output and the possible efforts that would be required to bring the segments to a post-edited version. By attributing an 'Excellent' score to an MT output, evaluators judge that the segment does not need to be modified. Asking evaluators to take post-editing modifications into account during the evaluation was deemed necessary to respond to the question raised by O'Brien (2006: 177):

> If an evaluator rates a sentence that has been machine translated as "excellent", does this mean that the sentence will not require post-editing?

In this study, the attribution of 'Excellent' scores may be compared with the validation of 100% TM matches. The suggested segment is read, compared with the original and validated. On the other hand, 'Medium' and 'Poor' scores will show that the MT output may be cumbersome for post-editors to correct. Based on Krings' findings (2001: 541), however, it will not be possible to assume that a 'Poor' segment would take longer to post-edit than a 'Medium' segment. In his study, Krings (ibid) found out that 'medium quality causes more difficult PE overall'.

At another level, these scores will also show how comprehensible the MT outputs would be for end-users. This is facilitated by the emphasis that was placed on accuracy, comprehensibility, and well-formedness of the output with the higher score ('Excellent'). In theory, 'Good' and 'Excellent' scores would be understood by end-users that have no access to the source text. In order to achieve this objective, evaluators were asked to read target and source segments in turn to try and view output from an end-users' perspective. These two situations are similar up to a certain point. Evaluators have access to the source text to check that they understood correctly all that they had read. Besides, the limitations of this artificial comparison are obvious, since evaluators are likely to find the MT output more comprehensible due their more in-depth knowledge about the topic than end-users. This is not surprising, as Danks and Griffin (1997: 172) note that 'if readers have knowledge that is specific to the text, then processing is facilitated.' Despite this limitation, it was essential to include this requirement with regard to end-users, as long as the evaluation was performed by evaluators who were familiar with the audience targeted by this type of technical content. Other evaluation requirements also had to be considered when choosing the evaluators. These requirements are presented in section 3.4.3.

3.4.3. Evaluation requirements

White (2003: 219) remarks that 'very ordinary things can affect someone's ability to be consistent in their judgments. Specifically, they will get tired, bored, hungry, or fed up with the process of evaluating.' This statement is particularly relevant, if the evaluation process is to be performed by people who are not focused on the task.

A choice had to be made with regard to the selection of evaluators. Should Symantec in-house translators and reviewers be used for the evaluation process, or should an external agency be hired? After considering the pros and cons of each alternative, it was decided to choose the former option. The main reason for selecting in-house translators and reviewers was their familiarity with the topic and Symantec-specific terminology. Besides, it was essential that both sets of MT outputs were scored by the same person within a defined timeframe. These parameters would be difficult to control if the evaluation task was outsourced. It was decided to use four evaluators for each language pair. This decision was in line with Dyson and Hannah's recommendation (1987: 166) to use no less than three evaluators, and Arnold et al.'s conviction that 'four scorers is the minimum; a bigger group would make the results more reliable' (1994: 162). For both German and French, two of the evaluators were translators with post-editing experience, while the other two were linguistic reviewers or experts without post-editing experience. Using evaluators that were familiar with a specific type of content also meant that they may be more stringent when judging the quality of the output.

As an addition to the metrics presented earlier, the evaluators were given the following extra guidelines. They were told to complete the evaluation of MT output A (304 examples for a total of 4456 source words) before starting the evaluation of MT output B (304 examples for a total of 4645 source words). This guideline was introduced to minimise the effect of reading an output for a second time, even if slightly different from the first initial one. White (2003: 218) warns that 'the second time they see the translation, they have an informed idea of what the expression is supposed to say, and this affects their judgment of whether it actually says it or not.' Section 4.2 of Chapter 4, which focuses on the consistency and the reliability of the results, will show whether this objective was achieved.

Evaluators were not told which inputs had been treated by CL rules. They also did not know whether the outputs had been mixed. It was a conscious decision not to mix the outputs at random since it was assumed that the time allocated to perform the evaluation (two weeks) and the size of the test suite would prevent evaluators from being influenced by their previous scores. The evaluation process can be regarded as a 'single-blind experiment' (Levine & Stephan, 2005: 6) because only

the researcher knew which segments were controlled. The purpose of this approach was to ensure that evaluators would not be influenced in any way, thus maximising the validity of the results presented and discussed in Chapter 4.

3.5. Summary

This chapter presented the environment that was used to carry out an evaluation of MT segments. Section 3.2 of this chapter focused on the selection of data used in the test suite environment that was specifically designed for this study. This environment contained two sets of segments that were extracted from the corpus based on their violations of specific CL rules. Section 3.3 of this chapter detailed the parameters used to machine-translate these segments. Finally, potential evaluation metrics were reviewed in the light of this study's objectives. Specific metrics were designed based on two widely-used MT evaluation criteria. Chapter 4 will present the analysis of the evaluation results.

Chapter 4

Chapter 4: Analysing the impact of individual CL rules at the segment level

4.1. Objectives of the present chapter

In this chapter, the results of the evaluation process described in Chapter 3 are analysed. To meet the first objective of this study, the analysis of these results will be used to determine whether certain CL rules are more effective than others in improving the comprehensibility of MT output at the segment level. Section 4.2 of this chapter presents the method used to analyse the results. In Section 4.3, each rule is reviewed based on its level of impact of the comprehensibility of MT segments. Finally, the main findings of this analysis are summarised in section 4.4.

4.2. Preparing for the analysis of the results

In order to analyse quantitatively the data collected from the evaluation process, it was decided to replace the scores given by the evaluators by numeric values. 'Excellent' was replaced by 4, 'Good' by 3, 'Medium' by 2, and 'Poor' by 1. These replacements were essential to be able to measure continuous variables, such as the median score of a given example, or the improvement of its quality thanks to the rewriting. For instance, if a German MT example A receives 3 M scores and one G score, its median score is 2. If the corresponding MT output B receives 2 'Good' scores and 2 'Excellent' scores, its median score is 3.5. The difference between the two median scores given to each example can then be calculated. In the example used earlier, a raw score of 1.5 is obtained when the median scores are calculated and subtracted from one another (3.5 − 2). The following formula shows how examples' raw scores are calculated for each target language:

Example's raw score = Example B's Median − Example A's Median

The decision to use median scores instead of average scores is made to ensure the consistency and reliability of the results. Section 4.2.1 discusses these issues by providing an overview of the results of the evaluation process.

4.2.1. Consistency and reliability of the results

As mentioned in Hovy et al. (2002: 6), 'an important matter is inter-evaluator agreement, reported on by most careful evaluations.' The next section focuses on the reliability of the evaluation results, based on the agreement among evaluators.

4.2.1.1. Reliability of the results

Before examining in detail the scores that have been given to MT outputs A and B, the level of general agreement among the evaluators of the same language pair is assessed. Four evaluators were used, so relative agreement is achieved when a majority of three evaluators gives similar, better, or worse scores to both outputs, regardless of the type and level of change. For instance, one German evaluator may decide to give 'Medium' scores to both MT output A and MT output B of a particular segment. In this case, no improvement is noted between the two segments. Likewise, a second evaluator may decide to give 'Good' scores to both outputs, confirming the lack of improvement. On the other hand, if the last two evaluators decide to give MT output A 'Medium' scores, and MT output B 'Good' scores, an improvement is noted. In this situation, however, a consensus is not reached since a majority of three evaluators is not obtained. This situation is referred to in this section as a disagreement. Table 4-1 shows the level of consensus among the evaluators based on the total number of segments (304):

Target Language	Disagreement	Agreement (Worse)	Agreement (Same)	Agreement (Better)
German	33% (99/304)	4% (11/304)	41% (126/304)	22% (68/304)
French	34% (103/304)	2% (6/304)	31% (93/304)	32% (102/304)

Table 4-1: Inter-evaluator agreement

These results show that at least three German evaluators agreed on 205 segments (67%) with regard to the general trend of change from MT output A to MT output B (same, better, or worse), while at least three French evaluators agreed on 201 segments (67%). Despite having the exact same instructions, these results demonstrate that evaluators did not interpret them in the same way, or decided to rate output differently.

This lack of systematic consensus is confirmed by the differences existing at the segment level between the scores provided by evaluators. These differences may not be the main focus of this study, but it is worth establishing how divided evaluators were when attributing scores to specific segments. Our data show that scores of {4,4,1,1} for a specific segment never occur for both sets of French and German MT outputs, but that scores of {1,2,3,4} appear twice (in the French MT output A data set).

Evaluators have hesitated between 'Good' and 'Medium' scores for a small number of segments. Scores of {3,3,2,2} appear in the data set, representing 8% of the German results (for both MT output A and MT output B), 7% of the French results for MT output A and 3% of the French results for MT output B. For these segments, a fifth evaluator may have helped establish a pattern.

Overall, these numbers confirm that the results of the evaluation are only congruent for about two-thirds of the segments in each language. When segments are analysed in detail in the next part of this chapter, the focus will be placed on the segments for which a consensus has been reached. After examining the general trend of agreement among evaluators, the consistency of the scores provided by the evaluators is scrutinised.

4.2.1.2. Consistency of the results

In some cases, certain MT output B examples were identical to MT output A examples, because both inputs were semantically identical and the CL rule had no impact on the comprehensibility of the MT output. It is therefore possible to check whether evaluators assigned consistently the same score to both MT outputs. For instance, Bernth (1998: 35) recommends that relative pronouns should not be omitted in front of past-participles. This rule (rule 9 in Appendix A) was evaluated with eight segments, by introducing a relative pronoun and a form of 'be' in MT output B. The replacement is shown as follows:

⊠ Repeat this step for each registry key listed in the error.

Ⓐ Wiederholen Sie diesen Jobstepp für jeden Registrierungsschlüssel, der im Fehler aufgeführt wird.

☑ Repeat this step for each registry key that is listed in the error.

ⓑ Wiederholen Sie diesen Jobstepp für jeden Registrierungsschlüssel, der im Fehler aufgeführt wird.

In the above example (example 41 in appendices F and G), German MT output A and MT output B are identical, but the scores given by German evaluators were not consistent across both versions ({2,3,3,3} for MT output A and {3,4,4,2} for MT output B). Such a situation was reproduced with 41 other segments in German, but only 29 segments in French, showing that rewritings had less impact on German output B than French output B.

Table 4-2 shows the number of identical segments in both language pairs and the consistency of each evaluator:

Target Language	Evaluator 1's consistency	Evaluator 2's consistency	Evaluator 3's consistency	Evaluator 4's consistency
German	88% (37/42)	64% (27/42)	81% (34/42)	64% (27/42)
French	40% (12/30)	60% (18/30)	73% (22/30)	73% (22/30)

Table 4-2: Evaluators' consistency on identical segments

These percentages show that the German evaluators were much more consistent with their own scoring, averaging 74.25% consistency on identical segments, as opposed to 61.5 % for the French evaluators, strongly affected by the 40% consistency of evaluator 1. This lack of consistency may be explained by the difficulty for human evaluators to maintain the same level of concentration when the evaluation task is repetitive, and requests a focused attention to detail. It also corroborates Krings' findings (2001: 11) that evaluators have a tendency to change internal criteria over time. It could be argued that asking them to score both segments in succession would have limited this consistency problem, since they could have compared both segments and adjusted their scores. However, the objective of this study was not to perform a comparison of both segments during the evaluation process itself, but rather during the analysis of the data collected during this evaluation.

Krings (ibid) also explains that in some cases evaluators and 'post-editors become so accustomed to the phrasing produced by the MT output that they will no longer notice when something is wrong with it'. In the present evaluation, the scoring variation was often performed in favour of MT output B. In the above example, three of the German evaluators thought that the above example had been improved by the rewriting, whereas one of them thought that it had become worse. However, the tendency to favour MT output B as opposed to MT output A was not as clear-cut in both languages. 46 cases of inconsistent scoring were noted across the four French evaluators, and in 80% of the cases, the variation was translated into an improvement. However, only 58% of the 43 cases of inconsistency in German were positive improvements. It should be mentioned that almost all of these negative or positive variations returned a difference of 1, whereby, for instance, an evaluator attributed a 'Medium' score to a segment A and a 'Good' score for an identical segment B. Only six cases of inconsistency showed a variation of 2 among the French evaluators, and two cases among the German evaluators. Since French evaluators proved less consistent and more inclined to favour MT output B than their German counterparts, a refined scoring mechanism must be used to assess the effectiveness of each rule. The selection of such a scoring mechanism is discussed in the next section, where the results of the evaluation are presented for each rule.

4.2.2. Selection of scoring mechanisms for the analysis of the results per CL rule

A combined approach is used to analyse the results of this evaluation so that the impact of CL rules on the comprehensibility of MT output can be assessed in a reliable manner. Based on the issues of consistency and reliability discussed in the previous sections, using the evaluators' average scores obtained for both versions of a segment is not an appropriate scoring mechanism. Such average scores would be impacted by unusual scores provided by individual evaluators for specific segments, be it out of tiredness, frustration, or keyboard errors. Median scores are therefore used for each segment.

4.2.2.1. Identifying a generic score for each CL rule

To justify this choice, the sets of scores obtained for example 41 in German are presented again: {2,3,3,3} for MT output A and {3,4,4,2} for MT output B. If average scores were calculated for each of these two sets of scores, 2.75 and 3.25 would be obtained for MT output A and MT output B respectively. The score 2.75 is misleading because it does not reflect the fact that most evaluators thought this segment was understandable despite the minor grammatical errors it contains. On the other hand, a median score of 3 provides an accurate picture of the consensus obtained with three of the evaluators. Such differences are essential to take into account when examining the amount of improvement provided by a CL rule's rewriting recommendation. The raw score of this example is then calculated by subtracting the median score of MT output A from that of MT output B, which yields 0.5 (3.5 − 3). Since the 54 CL rules have been evaluated with a different number of examples, their generic scores must be calculated by averaging all of their examples' raw scores, as shown by the following formula:

$$\text{Generic score} = \frac{\sum (\text{Example B's Median} - \text{Example A's Median})}{\text{Number of Examples}}$$

Separate generic scores are calculated per target language to compare and contrast the results obtained for each of the rules in French and German. These scores, however, are unlikely to be sufficient to draw final conclusions on the overall effectiveness of all rules. Our understanding of certain rules may have to be refined by looking at specific examples, before stating that the rule is always effective or ineffective. By using this approach alone, it would not be possible to find out whether certain examples within each rule would have provided better results, had they not been polluted by other violations. A micro-evaluation of rules providing mixed results seems necessary to avoid this issue. Van Slype (1979: 14) mentions that one of the methods used during a micro-evaluation of translation quality implies a diagnostic level, whereby an 'analysis of the causes of errors input, and an analysis of the source language' is performed. By performing an evaluation of the results at the example level, it should be possible to determine why a CL rule had a specific type of impact, be it positive, negative, or zero. It should then be possible to

determine whether the scope of a rule could be restricted by eliminating patterns that did not improve the MT output, or in certain cases made it worse.

4.2.2.2. Using 'extra' scores to refine the analysis

Since one of our objectives is also to isolate the rules that provide the most important improvements, a second set of scores is required for each rule. Knowing that a rule improves MT output is one thing, but knowing by how much it improves MT output is another. In order to provide answers to this question, examples that have been greatly improved by the CL rules' rewriting will be counted. When using this second approach, the scoring agreement discrepancies noted earlier must be taken into account to ensure that the rules' scores do not originate from inconsistent scoring.

The following restrictions will therefore apply for the generation of such 'extra' scores:

- Examples with identical MT output A and MT output B are excluded

- Examples for which a consensus was not reached by at least three evaluators are excluded. An improvement must be noted in the raw scores of three evaluators.

These 'extra' scores are further divided to take into account the level of improvement brought by the application of a given CL rule. Two levels of positive 'extra' scores are defined as follows:

- For positive 'extra' scores of level 1, the median score of an example's output A must be inferior to 3, and the median score of the corresponding output B must be superior or equal to 3. The CL rule's rewriting ensured a major improvement in the comprehensibility of the MT output, since a majority of evaluators rated the example's output A as 'Poor' or 'Medium' and the example's output B as 'Good' or 'Excellent'.

- For positive 'extra' scores of level 2, the median score of an example's output A must be equal to 3, and the median score of the corresponding output B must be equal to 4. The CL rule's rewriting

ensured a minor improvement in the comprehensibility of the MT output, since a majority of evaluators rated the example's output A as 'Good' and the example's output B as 'Excellent'.

Two levels of negative 'extra' scores are defined as follows:

- For negative 'extra' scores of level 1, the median score of an example's output A must be equal or superior to 3, and the median score of the corresponding output B must be inferior to 3. The CL rule's rewriting had a major negative impact on the comprehensibility of the MT output, since a majority of evaluators rated the example's output A as 'Good' or 'Excellent' and the example's output B as 'Medium' or 'Poor'.

- For negative 'extra' scores of level 2, the median score of an example's output A must be equal to 4, and the median score of the corresponding output B must be equal to 3. The CL rule's rewriting had a minor negative impact on the comprehensibility of the MT output, since a majority of evaluators rated the example's output A as 'Excellent' and the example's output B as 'Good'.

The remainder of this chapter concentrates on the distribution of CL rules based on the impact suggested by their scores, including their 'extra' scores.

4.2.2.3. Distribution of CL rules based on their scores

As explained in the previous section, generic scores for each CL rule are calculated by dividing the sum of each example's raw score by the number of examples selected for each rule. For instance, rule 48, *'Use only "because", never "since" in a subclause of reason'*, mentioned in Adriaens and Schreurs (1992: 598), was evaluated using four examples. A total of eight MT outputs was scored by four evaluators, so the results for this rule's first example look as follows in German: {2,2,2,1} for MT output A (with a median of 2), and {2,2,2,2} for MT output B (with a median of 2). This example's raw score was therefore zero. This zero figure was added to the rule's other three examples' raw scores (0, 0, and 0.5) to produce a generic score of 0.125 for the CL rule in German. On the other hand, the rule provided a generic score of 0.625 in French. The same approach was used for all rules, and a snapshot of these results is provided in Figure 4-1. The scatter plot

diagram is not used to show the evaluation of a variable over time, but rather to compare the distribution of the scores per CL rule in each target language.
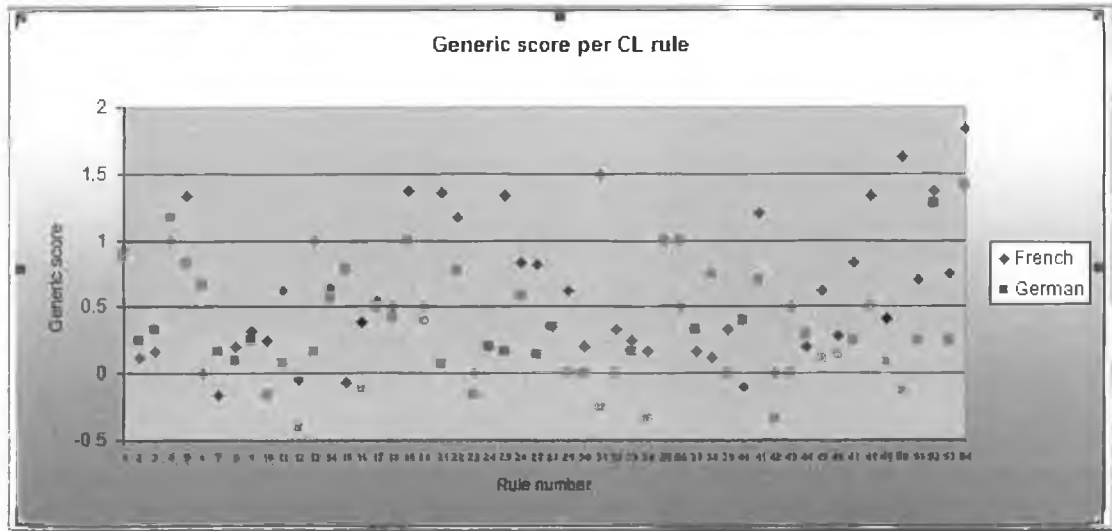

Figure 4-1: Generic score per CL rule

At first glance, CL rules appear to be more effective for French than for German. With these scores alone, it is, however, impossible to advance with certainty any explanations for this trend. The rules could be more effective in French due to the relative poor quality of French MT output A compared to that of German MT output A. They could also be more effective because of the inferior quality of German output B compared to that of French output B. A combination of both explanations is also possible. Furthermore, a rule may have a relatively high generic score, but fail to return any positive 'extra' scores. This was the case with rule 13, *'Do not use the impersonal passive voice'*. This rule obtained a generic score of 1 in French, but none of the three examples returned 'extra' scores that show a substantial improvement in the comprehensibility of the output. This discrepancy between generic and 'extra' scores suggests that the latter scores should be used as the main criterion to decide on the effectiveness of a particular rule. In order to classify rules, the following categories are going to be used:

- Rules that have a major positive impact on the comprehensibility of MT output in both languages

- Rules that have a limited positive impact on the comprehensibility of MT output in both languages

- Rules that have no impact in both languages

- Rules that have a negative impact in both languages

- Rules that mostly improve the comprehensibility of MT output in one language pair

After calculating 'extra' scores for each rule, rules were placed in one of these five categories based on 'frequency' scores that were calculated out of their French and German 'extra' scores. The following formula was used to calculate the 'frequency' score of each rule:

(((Number of positive 'extra' scores of category 1)*100)/Number of Examples) + (((Number of positive 'extra' scores of category 2)*100)/Number of Examples)/2 - (((Number of negative 'extra' scores of category 1)*100)/Number of Examples) - (((Number of negative 'extra' scores of category 2)*100)/Number of Examples)/2

This formula confirms that 'extra' scores of category 2 were not rewarded in the same manner as 'extra' scores of category 1, so as to reflect the level of comprehensibility improvement brought by the CL rule. Since the present study was the first to use this type of scoring mechanism, no pre-defined category was available to determine thresholds for rules with major positive impact or rules with limited impact. It was therefore decided to populate these five categories using arbitrary criteria, which are detailed in sections 4.3.1 to 4.3.6.

4.3. Analysis of the results per CL rule

4.3.1. CL rules that have a major positive impact in both languages

For a rule to be initially considered as having substantial impact on both French and German MT output, the following requirement had to be met: the frequency score of the rule must be greater than 33 for both French and German. As mentioned in the previous section, this threshold was chosen arbitrarily. This threshold, however, suggests that the rule is very effective at least every three segments. The 11 rules that met this requirement are discussed in turn to determine whether these rules should remain in this category. These 11 rules are listed in Table 4-3 in no particular order (the frequency scores have been rounded up):

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 54: Check the spelling.	6	67	67	4	0	4	0
Rule 52: Use single-word verbs.	11	64	54	6	2	6	1
Rule 4: Do not omit hyphens in Noun + Adjective or Noun + Past Participle structures.	3	33	33	1	0	1	0
Rule 19: Do not use pronouns that have no specific referent.	4	50	62	2	0	2	1
Rule 35: Avoid unusual punctuation.	2	50	50	1	0	1	0
Rule 5: Do not use more than 25 words per sentence.	3	33	100	1	0	3	0
Rule 1: Avoid ambiguous coordinations by repeating the head noun, or by changing the word order.	8	37.5	37.5	3	0	2	2
Rule 22: Do not coordinate verbs or verbal phrases when the verbs do not have the same transitivity.	9	56	56	4	2	3	4
Rule 41: Avoid embedded parenthetical expressions introduced by commas or dashes.	5	50	70	2	1	3	1
Rule 25: Two parts of a conjoined sentence should be of the same type.	3	33	67	1	0	2	0
Rule 48: Use a question mark only at the end of a direct question.	3	33	67	1	0	2	0

Table 4-3: Rules with major positive impact in both languages

Rule 54, '*Check the spelling*', showed significant improvements in both languages, since misspelt words such as 'Runing' or 'retarting' remained untranslated in the target outputs. Systran 5.0's auto-spelling correction function, however, correctly identified the spelling error in the following segment: 'This sould be "Permit All"'. Such results confirm that misspelt words affect the comprehensibility of the corresponding MT output. Eliminating such errors in the source content is therefore a priority, which will also improve the credibility of source documents. Hammerich and Harrison (2002: 115) mention that 'typos and other errors suggest that the content may also be erroneous'.

The results obtained for rule 52, '*Use single-word verbs*' confirmed that when a verb is used in conjunction with a particle, and this particle is analysed as a preposition, the source input may be parsed incorrectly. Incorrect parses will then greatly affect the comprehensibility of MT output, especially when the particle does not immediately follow the verb, as shown by the following example:

☒ This document describes how to turn Norton AntiVirus 2005 Internet Worm Protection on and off.

These results confirm the recommendations of Fuchs and Schwitter (1996: 127), Mitamura (1999: 47), and Bernth and Gdaniec (2001: 187). The scope of this rule is not limited to the use of verbs with particles. Rewritings addressing complex verbs, such as 'get to work' returned 'extra' scores of level 1 in both French and German output when these complex verbs were replaced with a simple verb such as 'use'.

One of the rules operating at the lexical level is rule 4, '*Do not omit hyphens in Noun + Adjective or Noun + Past Participle structures*'. The implementation of such a rule was justified Bernth and Gdaniec (2001: 193), who realised that it was 'quite common (at least in U.S. English) to omit the hyphen between a noun and a post-modifying past participle.' Mitamura (1999: 47) also noted that 'hyphenation should be consistently specified', which appeared to be a more appropriate angle to study the effectiveness of this rule, since post-modifiers can sometimes be adjectives instead of past participles, as in 'virus-free', or 'industry-standard'. This rule was evaluated with three examples and provided almost identical scores for French and German. It should be pointed out that the hyphenated words had been

coded as dictionary entries to study the impact of the slight typographical change introduced by the hyphen. The rule only proved effective for example 18, which contained an adjective as post-modifier, 'virus free'. The other two examples used (examples 19 and 20), were properly handled by the MT system, so the rewritings proved otiose. Based on these mixed results, it is difficult to generalise but the significant impact of the rewriting for example 18 (with two 'extra' scores of level 1) suggests that this rule is equally important for both French and German.

The next rule to provide indications of significant improvements in both French and German concerns a specific part of the rule prohibiting the use of pronouns. Rule 19, *'Do not use pronouns that have no specific referent',* echoes one of the COGRAM rules ("It" must refer to a preceding noun, Adriaens & Schreurs, 1995: 130). This rule was evaluated using four examples and the rewrites clearly improved the comprehensibility of both MT outputs. Pronouns with no specific referent, such as 'it' in 'makes it easy to...', were not handled correctly by the MT system in both languages. Referential ambiguity was also introduced in examples 114 and 117, in which pronouns were placed after a noun they did not refer to. From a CL implementation perspective, however, identifying these specific cases of referential ambiguity is bound to be a difficult task for a CL checker without any semantic analysis.

The description of rule 35 may sound very generic, especially as it was only evaluated with two examples: *'Avoid unusual punctuation'.* Two examples containing unusual punctuation usage were found in the corpus when extracting sentences for other rules. It was decided to test this rule as a separate rule to evaluate the impact of such violations. Each segment returned an 'extra' score of level 1 either for French or German. For instance, the original segment of example 203 contains dashes that are used to insert explanatory comments in the middle of a sentence: 'Type--or copy and paste--the following file names'. Since the dashes are not separated by spaces, an analysis problem occurred during the translation process, and the first word of the sentence was tokenised alongside the two following dashes. The effectiveness of the corresponding rewrite was, however, more significant in German and French due to unexpected noise created by the mistranslation of the homograph 'copy' in French. In the other example, where a comma had been

incorrectly placed before a noun, the exact same problem happened in German. The effectiveness of the rewrite was blurred by the mistranslation of the homograph 'copy' as a verb instead of a noun. This detailed analysis was helpful in identifying two clear cases of punctuation usage that should be avoided before an MT process.

Rule 5, which limits the number of words to 25, was evaluated with three examples only. It must be said that no standard rewrites exist for this rule, so the evaluation results depend heavily on the examples that are chosen to test this rule. One could even argue that such a rule is too vague to generate consistent improvements, but if no restriction is placed on sentence length, complex sentences such as the following may be found:

⊠ If you have Windows Me and you have run Windows System Restore and you see the error message "Unable to initialize the virus scanning engine" before you see error (3019,6), then follow the instructions in the document Error: "Unable to initialize virus scanning engine. . ." after running Windows System Restore or installing Norton AntiVirus 2003/2004.

This rule therefore showed clear signs of improvements in both French and German, with a common 'extra' score of level 1 obtained by example 21. It should be said, however, that not all sentences that are compliant with this rule will produce comprehensible output, especially if essential words are omitted due to conciseness. This issue will be further discussed when the rules governing the use of ellipses are discussed. Despite these limitations, reducing sentence length is crucial. This was confirmed by the fact that this rule was the only one in the set of 54 to return positive 'extra' scores for all three examples in French.

Signs of clear improvements were also obtained with rule 1, *'Avoid ambiguous coordinations by repeating the head noun, or by changing the word order'*. This rule, which builds on Systran's recommendation to write complete compounds when possible (Systran, 2005: 176), was evaluated with eight examples. Bernth (1998:33) also states that 'a real ambiguity occurs when a conjoined noun phrase premodifies a noun. In this case the scope of the premodification can be highly ambiguous'. This seems to be especially true when the coordination is punctuated with commas. Out of the eight examples used to test this rule, two returned 'extra' scores in both

French and German. Example 2 provided the best result for both languages, which returned an 'extra' score of level 1, since seven evaluators thought that the following example was improved by the rewrite:

- ☒ On the Schedule tab, fill in the Schedule Task, Start time, and Schedule Task Daily fields.
- A. Sur l'onglet Planification, complétez le temps de tâche, de Démarrer de programme, et les champs Tâche quotidienne planifiée.
- ☑ On the Schedule tab, fill in the Schedule Task field, the Start time field, and the Schedule Task Daily field.
- B Sur l'onglet Planification, complétez le champ Tâche planifiée, le champ Début, et le champ Tâche quotidienne planifiée.

In the original example, 'fields' is a head noun used at the end of the sentence to encompass the three option fields that must be completed by users. This type of structure may not be too ambiguous for a human reader, even though the end of the sentence has to be reached before it can be fully understood. However, the French output A shows that the source structure created analysis problems for the MT system. The translation 'champs' in MT output A was not recognised as being the head of several compounds, which was omitted twice to avoid repetitions. This decision was taken to make the ST as concise as possible. Conciseness, however often leads to ambiguity. If words are missing in the ST, MT systems must guess whether certain words ought to be restored in the TT. This issue is highlighted by Bernth and Gdaniec (2001: 207) who stress that the 'omission of information, be it ellipsis or missing sentence delimiters, causes the MT system to go into what (they) might call "guessing mode"'. As shown with the previous example, this strategy proves costly in specific cases. Not only is the syntax of the TL affected, but the French MT output A shows that certain user dictionary entries were not used during the translation process. Using elliptical structures may have the unfortunate consequence of negating dictionary work performed prior to the MT process.

It is a common dictionary practice to separate head nouns from modifiers. Petrits (2001: 9) mentions that the European Commission's dictionaries are divided into 'dictionaries for individual words and dictionaries for expressions'. This strategy, which avoids the duplication of dictionary work, can be implemented in Systran 5.0

with the use of lookup find operators (Systran, 2005: 136). For instance, a main user dictionary may contain an entry for the word 'field' preceded by a lookup operator. This lookup operator will point to any expression contained in the reference dictionary. 'Schedule Task', 'Start time', and 'Schedule Task Daily' would all be included in the reference dictionary. This technique has two advantages: Firstly, duplicate work is avoided, because there is no need to have both 'Schedule Task field' and 'Schedule Task' as separate entries. Secondly, the translation of the syntactic relationship between the head noun 'field' and its premodifiers can overwrite the MT system's default compounding rule which may not be appropriate in certain cases. For instance, the preposition 'de' would be added by default to link the head 'champ' with the postmodifier expression corresponding to 'Schedule Task'. In this particular example, this preposition is superfluous and would have to be removed during the post-editing process. Considering the frequency of these compounding structures in technical content in which software options are described, it seems cost-effective to avail of this referencing dictionary strategy. Typical nominal groups include words such as 'tab', 'menu', 'list', 'folder', or 'box', being preceded by user interface options such as 'Schedule Task', 'View', 'Save as', or 'Scan now'. These options are used as epithets to modify the head noun despite the fact that they contain verbs. Bloor & Bloor (2004: 139) mention that 'the function of Modifier can be realised by various word classes, most frequently by determiners, numerals and adjectives as Premodifiers'. When it comes to technical support documentation, and more generally software documentation, verbs must be added to this list, as shown in the above example with 'Schedule Task'.

Rule 22 deals with the use of coordinated verbs that do not share the same valency. One of them may be used in a direct transitive way, when the other is used indirectly with a preposition. This rule should therefore be regarded as a specific part of the rule suggesting that coordinated verbs should be avoided (Mitamura, 1999: 47). Coordinations are indeed negative translatability indicators used in the Logos translatability index (Gdaniec, 1994). Rule 22 returned a very high number of 'extra' scores of level 1 (three for French and four for German) and level 2 (four for French and for German), showing that the comprehensibility of MT output can be significantly improved when the object is inserted after the first verb, and a

pronoun is used after the second. The following example (example 133) returned 'extra' scores of level 2 in both French and German:

⊠ Scroll to and uncheck the following entry.

A Allez à et déselectionnez l'entrée suivante.

A Blättern Sie zu und deaktivieren Sie den folgenden Eintrag.

☑ Scroll to the following entry and uncheck it.

B Allez à l'entrée suivante et déselectionnez-la.

B Blättern Sie zum folgenden Eintrag und deaktivieren Sie ihn.

The results obtained for rule 22 must be contrasted with those obtained for rule 23. Rule 23 was evaluated with 3 examples, which did not return any 'extra' scores. The rewritings were ineffective because the original sentences produced perfectly comprehensible translations. Based on these results, it is possible to conclude that the scope of the very restrictive original rule concerning the use of coordinated verbs should be reduced to that of rule 22. This recommendation should prove effective as long as target languages are able to translate source transitive verbs in a simple transitive way, without resorting to phrasal verbs. An example of a potential problem is found in example 72 in French with the translation of the verb 'bookmark'.

Rule 41, *'Avoid embedded parenthetical expressions introduced by commas or dashes',* builds on Systran's recommendation to 'avoid multiple stacking' (Systran, 2005: 174). Dependent clauses are also used by Gdaniec (1994) as negative translatability indicators. While this rule may be regarded as an MT system-specific rule, its rewritings attempt to separate long sentences into shorter ones. This type of approach is based on the reuse philosophy discussed in section 1.4.2.2 of Chapter 1, and exemplified by example 227:

⊠ If you cannot change the drive letter, as a workaround to the problem, you can use the Windows Subst.exe utility to temporarily designate a folder as the c: drive.

☑ You cannot change the drive letter. As a workaround to this problem, you can use the Windows Subst.exe utility to temporarily designate a folder as the c: drive.

The use of two simple segments in the controlled example ensured that 'extra' scores of level 1 were obtained in both French and German. Breaking sentences into two segments was not always possible, but a reordering of the original sentences' elements improved the comprehensibility of the corresponding MT outputs. This was the case with example 230:

☒ A clean boot is similar to, but more thorough than, closing all applications.

☑ A clean boot is similar to closing all applications but it is more thorough.

While the translations of the controlled example were not perfect, their comprehensibility was improved with the removal of the interpolated clause and the introduction of a pronoun. This example suggests once more that the use of pronouns does not necessarily have a negative impact on the quality of the MT outputs as long as they have clear referents.

The effectiveness of rule 25, *'Two parts of a conjoined sentence should be of the same type'*, was evaluated with three examples. Two positive 'extra' scores of level 1 were obtained for French, but only one was obtained for German. The results obtained for this rule suggests that improvements can have positive impact in both languages.

Rule 48, *'Use a question mark only at the end of a direct question'* is explicitly mentioned by Adriaens and Schreurs (1992: 595). Such a rule is justified by the fact that the overall analysis of source input can be affected if the MT system attributes end-of-sentence properties to question marks appearing in the middle of segments. One can easily understand why technical writers may decide to include punctuation marks in the middle of sentences; they want their description to match exactly the content of the User Interface (UI), whose menus may contain direct questions, as shown with the following example:

☒ In the "What do you want to call this rule?" field, type Messenger Service and click Next.

A In dem „Wie möchten Sie diese Regel nennen??" Feld, Typ Messenger Service und klicken Sie auf Weiter.

- ☑ In the "What do you want to call this rule" field, type Messenger Service, and click Next.
- ฿ Im „Wie möchten Sie diese Regel nennen?" Feld geben Sie Messenger Service ein und klicken Sie auf Weiter.

In this example (example 265), three out of four German evaluators thought that the comprehensibility of the MT output was improved to such an extent that it would become understandable for users. This was probably because the sentence was analysed as one segment instead of two when the question mark was removed. It is interesting to notice that question marks were automatically inserted in both German outputs. This, however, was not the case in the French outputs, which suggests that different pre- or post-processing rules may be used by the MT system depending on the target language. This assumption cannot be verified without any access to the system's internal rules. In French, the effectiveness of the rule was not visible, since MT output B was affected by the mistranslation of the verb 'type', which was analysed and translated as a noun. Resolving ambiguity issues created by homographs will have to be taken into account in the next part of the study.

This rule was the last rule classified as having a major positive impact in both languages. Section 4.3.2 discusses the rules that have a limited impact on French and German outputs.

4.3.2. CL rules that have a limited positive impact in both languages

In order to qualify as a CL rule with a limited positive impact on both French and German outputs, the following requirement had to be met: the frequency score of the rule is greater than 0 in both languages but less than 33. The threshold for this frequency score was selected to reflect the fact that the rule returns 'extra' scores infrequently. Table 4-4 shows the 11 rules that fell into this category (in no particular order).

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 53: Avoid ungrammatical constructions.	4	12.5	25	0	1	0	2
Rule 36: The subject of a non-finite clause must be the same as that of the main clause.	4	25	25	1	0	1	0
Rule 14: Do not make noun clusters of more than three nouns.	7	29	14	2	0	1	0
Rule 17: Do not use a pronoun when the pronoun does not refer to the noun it immediately follows.	9	17	22	1	1	1	2
Rule 18: Do use pronouns, even when the pronoun immediately follows the noun it refers to.	7	21	14	1	1	0	2
Rule 51: Avoid -ing words.	8	25	25	2	0	1	2
Rule 20: Do not use an ambiguous form of 'could'.	5	10	20	0	1	1	0
Rule 28: Omit unnecessary words.	7	14	14	1	0	0	2
Rule 9: Do not omit relative pronouns and a form of the verb 'be' in a relative clause. The post-modifier is a past-participle.	8	12.5	12.5	1	0	1	0
Rule 2: Do not use ambiguous attachments of non-finite clauses (present-participles).	4	25	25	1	0	1	0

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 29: Place a purpose clause before the main clause.	4	12.5	25	0	1	1	0

Table 4-4: List of rules with limited positive impact in both languages

Rule 53, which proscribes the use of ungrammatical constructions, returned overall results that were similar in both languages. This rule was evaluated in the following scenarios:

- the misuse of the possessive 'its' or genitive marker 's'. Two examples were found in the corpus: 'Cannot locate ccSetMgr.exe or one of **it's** components' and 'the programs technical support'

- the lack of subject/verb agreement, an example being 'These conflicts appears to only happen when the computer is starting.'

The results suggest that the introduction of this rule had more impact on the well-formedness of the output than on its comprehensibility. One of the examples used to test the effectiveness of proper agreement between subject and verb even did not have any effect on German output, whereas it did for French (example 296):

☒ WinDoctor and One Button Checkup follow strict guidelines for what they considers valid or invalid.

A WinDoctor und One Button Checkup folgen strengen Korrekturlinien für, was sie gültig oder ungültig betrachten.

A WinDoctor et One Button Checkup suivent les directives strictes pour ce qu'ils considère valide ou non-valide.

☑ WinDoctor and One Button Checkup follow strict guidelines for what they consider valid or invalid.

B WinDoctor und One Button Checkup folgen strengen Korrekturlinien für, was sie gültig oder ungültig betrachten.

B WinDoctor et One Button Checkup suivent les directives strictes pour ce qu'ils considèrent valide ou non-valide.

As discussed in section 3.3.1.1, it is impossible to explain why a common source analysis problem produced different outputs without having access to the rules of the MT system.

Rule 36, *'The subject of a non-finite clause must be the same as that of the main clause'*, was evaluated with four examples. This rule is based on the grammaticality issue raised by Bernth and Gdaniec (2001: 178). Only one example returned an 'extra' score of level 1 in each target language. For French, it was example 204 whose non-finite clause was introduced by a past participle. For German, it was example 206, whose non-finite clause was introduced by an –ing word. A careful analysis of these four examples revealed that linguistic noise masked the effectiveness of the rule in German (for example, 'the dialog box graphic moves' was mistranslated in example 205). In three of the German outputs A (examples 204, 205, and 206), a personal pronoun was inserted in the clause preceding the main clause based on the subject of the main clause. This insertion did not appear in French outputs A since an equivalent non-finite clause can be also be used ambiguously in French. The challenge in preserving ambiguity in target languages has been described by Arnold (2003: 125), who believes that source ambiguity cannot be ignored in the same way in all languages. Based on this micro-analysis of these four segments, it can be concluded that this rule is more effective for German than for French when the subject of a non-finite clause introduced by an –ing word is not the same as the subject of the following main clause. When the non-finite clause is introduced by a past participle, the rule is also effective for the English-French language pair.

Rule 14, *'Do not make noun clusters of more than three nouns'* was evaluated with seven examples and returned fairly identical generic scores for French and German (0.6). This rule is often used in CL rule sets because of the potential difficulty for an MT system (and human readers) to determine the grammatical relationships existing between the elements of a long string of nouns. When connecting words are missing between nouns, ambiguity may be introduced by homographs. This is shown by the pre-CL segment of example 79:

⊠   Norton Internet Worm Protection features

In this pre-CL segment, 'features' could be analysed as a noun or as a verb by an MT system. In this particular case, the noun form was selected by the MT system so the introduction of the rule did not have any effect:

A   Fonctionnalités de Norton Internet Worm Protection
☑   Features of Norton Internet Worm Protection
B   Fonctionnalités de Norton Internet Worm Protection

Several types of rewritings exist for this particular rule, depending on the grammatical relationships of the elements of the cluster. For instance, two types of genitive structures can be used, either the Saxon genitive (example 81) or the Norman genitive (example 79). Other prepositions are sometimes required, as shown with the introduction of 'for' in example 76. The results obtained for this rule are therefore limited to the examples selected in this test suite. Making generalisations for this rule is difficult because ambiguous prepositions, such as 'against' or 'to', may sometimes be used to split clusters. These prepositions may be translated differently depending on the context in the target language, so their introduction may produce an MT output that is not necessarily better than the default output produced by a given MT system. Example 78 suggests that the system's default compounding mechanism was more effective in French than in German. The rule introduced in this example had less impact on the comprehensibility of the French output than the German output. This example returned an 'extra' score of level 1 in German, whereas little improvement was visible in French.

Rule 17, *'Do not use a pronoun when the pronoun does not refer to the noun it immediately follows'* was evaluated with 9 examples, and returned similar generic scores in French and German. The number of 'extra' scores was also fairly consistent across the two languages. Example 99 is worth discussing because it contains a case of referential ambiguity that was well-handled by the MT system.

☒   If an infection is found, you will be sent to a tool that removes it.

The pronoun 'it' contained in the pre-CL segment does not refer to the preceding noun, but the correct translation of the pronoun was produced by the MT system in both French and German. However, both groups of evaluators thought that the MT

output was more comprehensible when a full form of the noun was repeated in the translation. This result suggests that the well-formedness of a sentence does not ensure its comprehensibility. Overall, one cannot conclude that the effectiveness of this rule is visible on a consistent basis for the two languages evaluated. In certain cases, the MT system performed the right disambiguation, and in others, the ambiguity was acceptable in the target language. Such results were also noted with rule 18, *'Do not use pronouns, even when the pronoun immediately follows the noun it refers to'*. Since the style of the source text may become very repetitive and verbose with the re-introduction of nouns, the application of such rules will depend on the environment in which the CL is deployed.

A similar conclusion was also obtained when examining the results obtained for rule 51, *'Avoid -ing words'*. Due to the extreme ambiguity associated with –ing words, it was decided to evaluate the effectiveness of this rule using 8 examples. The scope of this rule had to differ from rules 2, 3, 38, and 46. The French generic score was higher than the German generic score, but 'extra' scores were more consistent across both languages, suggesting that the effectiveness of this rule did not favour a particular target language. As shown by example 277 in French, using two –ing words in succession can affect the comprehensibility of the MT output. This example was based on a title found in the corpus, 'Preparing for installing Norton SystemWorks 2005'. 'Ing' words are, however, very often used in headings, so implementing this rule in a systematic way would impact the acceptability of the source text. One possible solution to this problem may be to limit the number of 'ing' words to one per sentence, or to ensure that they do not occur after certain prepositions (as shown by the effective rewritings for examples 281 and 282 for French). In such cases, however, the rewrite must be handled correctly by the MT system, which was not the case with example 283, whose post-CL segment was more complex than the pre-CL segment. Specific 'ing' words may also be targeted, such as 'following' when it is synonymous with 'after' (as shown by example 279 which returned an 'extra' score of level 1 in German). Using such an approach will be considered in the next part of this study.

Rule 20, *'Do not use an ambiguous form of "could"'*, was evaluated with five examples, whose pre-CL segments contained a form of 'could' that was used with a

hypothetical meaning. Only two of these examples returned positive 'extra' scores in either French or German (examples 121 and 119 respectively). A micro analysis of the segments showed that past forms of 'can' were generated more often in German MT output A ('konnte' or 'konnten' instead of the conditional form, 'könnte' or 'könnten') than in French. It is, however, impossible to determine why these differences occurred (in example 120, for instance).

Rule 28, *'Omit unnecessary words'* must not be confused with rules 26 or 27. While the rules restricting ellipses attempt to preserve key words, rule 28 attempts to rid source text of redundant or tautological words. This rule was evaluated with seven examples and returned consistent generic scores in French and German. With the removal of single words such as 'just' or 'basic', or longer expressions such as 'and so on' which are void of any real meaning, the comprehensibility of MT output was improved. It is crucial to find the right balance between keeping words that will help the MT system disambiguate source input, and removing those that will impact negatively on the MT output.

Rule 9, *'Do not omit relative pronouns and a form of the verb "be"* in a relative clause when the post-modifier is a past-participle.' This rule was evaluated with 8 examples but only one example per language showed signs of effectiveness, example 40 for German, and example 42 for French. Making generalisations for this rule is difficult but it is possible to state that the polysemous nature of the verbs used in these sentences, 'caused' and 'made', may have provoked analysis problems for the MT system with the pre-CL segments.

During the evaluation of rule 2, *'Do not use ambiguous attachments of non-finite clauses (present-participles)'*, only one 'extra' score of level 1 was returned in both languages with example 12. After performing a micro-analysis of the four examples with which the rule was evaluated, it appeared that the improvement introduced by the rule in this example could have been influenced by two factors. It could have been influenced by the verb that was present in the non-finite clause. But it also could have been influenced by the type of clause that the ambiguous –ing word was introducing. Three examples contained a clause of manner introduced by the possibly ambiguous verb 'using'. This verb was parsed and translated properly in all

three cases. On the other hand, all evaluators agreed that the rewriting introduced in example 12 improved the comprehensibility of the MT output. The pre-CL segment of this example was different from the other three, since it contained an –ing word introducing a non-finite purpose clause. This result confirms that the choice of examples used to evaluate the effectiveness of CL rules can have some impact on the final formalisation of a rule. If exceptions are not added to a rule when such a rule is formalised, the rule may be used more often than it should be.

The last rule to show limited impact in both languages is rule 29, *'Place a purpose clause before the main clause'*. This rule was evaluated with four examples that contained purpose clauses introduced by infinitive verbs. These clauses were moved into first position in the post-CL segments. Only one example showed clear signs of improvement in both languages when the rule was applied, example 179:

☒ Follow the instructions in this section to repair the Intrusion Detection files.

A Suivez les instructions dans cette section de réparer les fichiers de détection d'intrusion.

A Befolgen Sie die Anweisungen in diesem Abschnitt, die Intrusion Detection-Dateien zu reparieren.

☑ To repair the Intrusion Detection files, follow the instructions in this section.

B Pour réparer les fichiers de détection d'intrusion, suivez les instructions dans cette section.

B Um die Intrusion Detection-Dateien zu reparieren, befolgen Sie die Anweisungen in diesem Abschnitt.

In this example, the preposition 'to' introducing the infinitive clause was mistranslated in the French MT output A and left untranslated in the German output A. This was not the case with the three other examples. These results show that the MT output was not consistently improved by the rule, but it is difficult to explain why results are not homogeneous. In future studies, it could be interesting to evaluate the impact of this rule on a language such as Chinese, whose clause order is much stricter than that of French and German.

4.3.3. CL rules that have no impact in both languages

Certain CL rules were classified as having no impact in both languages when their frequency score was equal to 0. In some cases, certain rules returned a positive 'extra' score which was cancelled out by a negative 'extra' score. These 10 rules are presented in Table 4-5:

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 7: Do not omit relative pronouns and a form of the verb 'be' in a relative clause. The post-modifier is an adjective.	3	0	0	0	0	0	0
Rule 23: Do not coordinate verbs or verbal phrases that share the same object	3	0	0	0	0	0	0
Rule 32: Start new sentences instead of using semi-colons.	3	0	0	0	0	0	0
Rule 34: Use the serial comma	3	0	0	0	0	0	0
Rule 39: Always write "in order to" before an infinitive in a purpose clause instead of just "to".	6	0	0	0	0	0	0
Rule 43: Avoid using "(s)" to indicate plural.	4	0	0	0	0	0	0
Rule 46: Do not use a participle to introduce an adverbial clause.	4	0	0	0	0	0	0
Rule 37: Ensure that relative pronouns or complex relative pronouns immediately follow their antecedent.	3	0	0	0	0	0	0

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 10: Do not enclose relative clauses, parenthetical remarks, or extraneous comments in parentheses.	6	0	0	1	0	0	1
Rule 30: Words such as "all", "one", or "none", may not appear alone.	5	0	0	0	0	1	0

Table 4-5: Rules with no impact in both languages

Rule 7, *'Do not omit relative pronouns and a form of the verb "be" in a relative clause, when the post-modifier is an adjective'*, was evaluated with three examples that did not show any signs of improvement. The lack of impact in French can be explained by the fact that relative pronouns can also be optional in this language. This was confirmed by the consensus obtained from French evaluators who attributing 'Excellent' scores to the three MT outputs A. The lack of impact in German can be explained by the fact that relative pronouns were automatically inserted in two of the three MT outputs A. These findings suggest that this rule is not effective when the post-modifier immediately follows the noun it qualifies. Results may be different if ambiguous partitive structures were used in front of post-nominal modifiers.

The impact of rule 23 was covered in section 4.3.1. The next two rules with no apparent impact in both languages concern two punctuation rules, rule 32 and rule 34. A micro-analysis of these examples confirmed the absence of positive impact in both languages.

Rule 39, *'Always write "in order to" before an infinitive in a purpose clause instead of just "to".* This rule, mentioned by Bernth and Gdaniec (2001: 187) and Systran (2004: 172) was evaluated with six examples. The French generic score was superior to the German generic score, due to the introduction of the more idiomatic translation 'afin de' instead of 'pour' in MT output B. This improvement, however,

does not appear valuable enough to justify the introduction of two extra words in the source text on a systematic basis. This finding echoes that of O'Brien (2006: 174), who found that the presence of such a negative translatability indicator ('to' instead of 'in order to') did not increase the post-editing effort. This should be good news to Ó Broin (2005: 37), who states that the elimination of these two extra words can save 40% of the translation cost for sentences containing five words. When reviewing other examples in the test suite, however, two mistranslations of the preposition 'to' introducing purpose clauses were found in the French output (examples 103 and 179). The typical infinitive structure was also missing from the German segments for these examples.

The introduction of rule 43, *'Avoid using "(s)" to indicate plural'*, did not improve the comprehensibility of MT output based on the results obtained in both languages. This can be explained by the fact that the MT system recognised this marker and automatically replaced it with a plural form in pre-CL segments 237 and 239. However, a singular form was generated in the translations of pre-CL segment 238, thus affecting the accuracy of the translation.

Overall, these results contrast with O'Brien's findings (2006: 177), which showed that this negative translatability indicator led to high post-editing effort. These different findings may be explained by several reasons. First of all, this linguistic phenomenon may have been handled in a different manner by the MT system used by O'Brien (IBM WebSphere). Second, the methodology she used to evaluate the impact of this negative translatability indicator on post-editing effort fundamentally differs from the approach taken in the present study. Her subjects had to *post-edit* MT output, whereas the evaluators used in the present study had to *rate* MT output by taking into account hypothetical post-editing requirements. These two different approaches appear to have produced different results.

Rule 37, *'Ensure that relative pronouns or complex relative pronouns immediately follow their antecedent'* was evaluated with three examples, but none returned any extra positive score. The generic scores were also low in both languages. A micro-analysis of the segments was performed to determine whether the effectiveness of this rule may have been blurred by other linguistic phenomena. This was the case in example 209, where the replacement introduced in the post-CL segment resolved

the initial issue, but introduced referential ambiguity. In the two other examples, the word 'that' did not introduce a genuine relative clause, but contained a nominal projection. This type of structure is related to some of the structures covered by rule 6. In the pre-CL segments of examples 208 and 210, however, the word 'that' was ambiguous because it was recognised as a relative pronoun by the MT system. Despite leading to ungrammatical MT output A, the evaluators' scores do not suggest any improvement with the rewriting. This may be partly explained by inconsistent scoring. For instance, the French evaluators divided their four scores into 'Medium' and 'Good' scores for MT output A of example 210, which did not lead to any improvement using the scoring mechanism described in section 4.2.2.2.

Rule 10, *'Do not enclose relative clauses, parenthetical remarks, or extraneous comments in parentheses',* differs from rule 42, since it does not allow for any material within parentheses. Rule 42 only proscribes parenthetical expressions that break the syntactic flow of a sentence. The effectiveness of rule 10 was evaluated using six examples, which produced different results depending on the target language. Before examining examples that returned negative scores, the example that returned positive 'extra' scores is discussed since it reveals source analysis differences. Taking this factor into account is important because it has an impact on the generalisations that can be made with regard to the effectiveness of certain CL rules. The parenthetical expression in example 48 seems to have been analysed differently by the MT system, since it produced different outputs:

- ⊠ Predefined system files (must be placed within the first 2 gigabytes of the drive)
- A Fichiers système prédéfinis (doit être placé dans les 2 premiers gigaoctets du disque)
- A Vorbestimmte Systemdateien (müssen innerhalb der ersten 2 Gigabytes des Laufwerks platziert werden)

The French output shows that there is no agreement between the initial verb of the parenthetical expression ('doit') and the subject of the preceding segment, 'fichiers'. This suggests that the parenthetical expression has been parsed without taking into account the first part of the segment. This is not the case in German output A, since the verb 'müssen' agrees with 'Systemdateien'. Nonetheless, a positive 'extra' score

of level 1 was obtained for this segment in German, while a positive 'extra' score of level 2 was obtained in French.

On the other hand, negative 'extra' scores were returned for two examples: 43 in French and 44 in German. In these examples, the pre-CL segments contained a parenthetical expression located at the end of the segment. The parenthetical expressions were therefore well separated from the rest of the segments. In example 43, ambiguity was unexpectedly introduced in the post-CL segment, as shown by the French example, which received a negative 'extra' score of level 2:

☒ Determine the address of the listserv (usually similar to listserv@symantec.com, or techsupp@symantec.com).

☒ Déterminez l'adresse du listserv (habituellement semblable à listserv@symantec.com, ou à techsupp@symantec.com).

☑ Determine the address of the listserv, which is usually similar to listserv@symantec.com, or techsupp@symantec.com.

B Déterminez l'adresse du listserv, qui est habituellement semblable à listserv@symantec.com, ou de techsupp@symantec.com.

The incorrect use of the preposition 'de' in front of 'techsupp@symantec.com' in MT output B shows that the comma preceding 'or' should have been removed from the post-CL segment. This would have prevented its incorrect analysis by the MT system.

A more significant problem was noted with example 44. The integration of the parenthetical comment into the main segment introduced complexity in the post-CL segment, which affected the comprehensibility of the German MT output. This finding suggests that the effectiveness of this rule is very limited. This was confirmed by examining examples containing parenthetical expressions that were embedded in the middle of the segments (in examples 45 and 46). Removing the parentheses in the post-CL segments did not improve the comprehensibility of the MT output in both French and German.

The effectiveness of rule 30, *'Words such as "all", "one", or "none", may not appear alone'* was evaluated using five examples. Only one of them returned a positive 'extra' score of level 1 in French (example 181), when "one" followed a preposition ("with"). One negative 'extra' score of level 1 was obtained with example 185, despite "none" being used on its own in the pre-CL segment. These results suggest that the effectiveness of this rule is not consistent, and that its impact minimal. In section 4.3.4, the CL rules that show signs of negative impact in both languages are discussed.

4.3.4. CL rules that show signs of negative impact in both languages

Only two rules fall into this category, even though some of the examples used for their evaluation returned positive 'extra' scores. The negative scores are presented in Table 4-6 with a "(-)" marker in the first row.

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (-)	German 'extra' score 2 (-)	French 'extra' score 1 (-)	French 'extra' score 2 (-)
Rule 42: Do not include parenthesized expressions in a segment unless the segment is still valid syntactically when you remove the parentheses while leaving the parenthesized expressions.	6	-33	0	2	0	1	0
Rule 12: Use the active voice when you know who or what did the action.	12	-30	4	3	1	0	2

Table 4-6: Rules that show signs of negative impact in both languages

The effectiveness of rule 42, *'Do not include parenthesized expressions in a segment unless the segment is still valid syntactically when you remove the parentheses while leaving the parenthesized expressions'*, was evaluated with six examples. The scope of this rule is different from that of rule 10. Whereas rule 10

proscribes any parenthetical expressions occurring in the middle of sentences, this rule focuses on parenthetical expressions that break the syntactic flow of the sentence. This difference is exemplified with two examples:

- ⊠ Log on as the user who saw the error (and who now has administrator access), and continue to the next section. (example 45 used for rule 10)
- ⊠ The (3021,2) error message prevents you from seeing the Norton AntiVirus interface. (example 236 used for rule 42).

Rule 42 returned one positive 'extra' score in French for example 45, but none for German. This may be explained by a different handling of punctuation marks in the two language pairs, as shown by both MT outputs A:

- A Die (3021.2) Fehlermeldung hindert Sie am Sehen der Norton AntiVirus Schnittstelle.
- A (Les 3021.2) messages d'erreur vous empêchent de voir l'interface de Norton AntiVirus.

The parenthetical expression was correctly analysed in the English-German language pair, since the translation follows the word order of the pre-CL segment. This was not the case with the English-French language pair.

The second rule to return negative 'extra' scores in both languages in a consistent manner is rule 12, *'Use the active voice when you know who or what did the action'*. This rule was evaluated using a large number of examples to diversify the types of agentive structures present in pre-CL segments. The examples returning negative 'extra' scores suggest that some of the rewritings proved more ambiguous than the original sentences. This is shown by example 70,

- ⊠ You want to know the most common load points that are used by viruses.
- A Sie möchten die meisten gemeinsamen Ladepunkte kennen, die durch Viren benutzt werden.
- ☑ You want to know the most common load points that viruses use.
- B Sie möchten die meisten gemeinsamen Ladepunkte kennen dass Virusgebrauch.

This example suggests that the use of the active voice in short relative clauses can introduce ambiguity. This finding was only partly confirmed with example 68, since it returned a negative 'extra' score of level 1 in German and a positive 'extra' score of level 1 in French.

Apart from examples 187 and 236, this is the only example in the test suite that produced opposite results in both languages. Example 68 is shown below:

- ⊠ Select other physical disk drives that you want to be protected by Norton GoBack.
- A Choisissez d'autres disques physiques que vous voulez être protégé par Norton GoBack.
- A Wählen Sie andere physikalischen Festplatten aus, die Sie von Norton GoBack geschützt werden möchten.
- ☑ Select other physical disk drives that you want Norton GoBack to protect.
- B Choisissez d'autres disques physiques que vous voulez que Norton GoBack protège.
- B Wählen Sie andere physikalischen Festplatten aus, dass Sie Norton GoBack schützen wünschen.

German output B shows that the relative pronoun in the post-CL segment was analysed as a conjunction with the introduction of the active voice in the relative clause. This does not seem to have been the case in French, even though 'que' is also ambiguous in that language. The comprehensibility of the French output was therefore improved by the introduction of the CL. Apart from example 64, this example is the only one that returned positive scores across both languages, so the effectiveness of the rule appears very limited overall. Such a finding goes against the recommendations present in the literature (Bernth & Gdaniec, 2001: 190, Wojcik, 1998: 118, Adriaens & Macken, 1995: 130). This also suggests that further evaluation is required to determine whether the scope of this rule should be reduced. It would also be very interesting to study the effectiveness of this rule in other languages, such as Chinese, due to the less common and more restricted use of the passive in this language (Ross, 2004: 241). For instance, Cardey et al. (2004: 40) had decided to exclude this structure from the Controlled English rule set they designed for translating medical protocols into Chinese.

4.3.5. CL rules that are more effective for French than for German

Based on the generation of 'extra' scores and their associated frequency score, certain rules appeared to be more effective in French than in German. Rules were initially placed in this category when one of two criteria was met:

- The rule's frequency score for French was greater than or equal to 33 and the rule's frequency score for German was less than 33.

- The rule's frequency score for French was greater than 0 and the rule's frequency score for German was less than or equal to 0.

One exception to these two criteria concerns rule 13, whose French frequency score was 0, but whose generic score was 1 (compared to a generic score of 0 in German). This suggested that the rule was more effective for French than for German. The 12 rules corresponding to the above criteria are listed in Table 4-7:

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (-)	German 'extra' score 2 (-)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 49: Do not use "this", "that", "these", or "those" on their own	6	0	17	0	0	1	0
Rule 45: Use only "because", never "since" in a subclause of reason	4	0	25	0	0	0	2
Rule 33: Use a comma before a coordinated clause.	6	0	17	0	0	1	0
Rule 21: Do not use the slash to list lexical items.	7	14	64	0	0	4	1
Rule 11: When appropriate, use an article (a, an, the) before a noun.	12	0	21	0	0	2	1
Rule 13: Do not use the impersonal passive voice.	3	0	0	0	0	0	0

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (-)	German 'extra' score 2 (-)	French 'extra' score 1 (+)	French 'extra' score 2 (+)
Rule 16: Avoid the use of progressive verb participles.	9	-8	6	1	0	0	1
Rule 50: Keep both parts of a verb together	4	0	50	1	0	2	0
Rule 31: Whenever possible, the use of "wh-" questions should be avoided	4	-12.5	75	0	1	3	0
Rule 26: Avoid the ellipsis of verb.	6	16	50	1	0	2	2
Rule 27: Avoid the ellipsis of subject.	14	7	43	1	0	6	0
Rule 47: Move document names, error messages, or section titles to independent segments.	6	17	67	1	0	4	0

Table 4-7: Rules with more positive impact in French than in German

The effectiveness of rule 49, *'Do not use "this", "that", "these", or "those" on their own'* was evaluated using six examples. Only one example returned a positive 'extra' score of level 1 in French (example 266). The pre-CL segment of this example contained a deictic pronoun introducing referential ambiguity. In German, the improvement brought by the application of the rule was masked by the complexity of the example. Based on these results, the effectiveness of this rule appears very limited in both languages.

Rule 45, *'Use only "because", never "since" in a subclause of reason'* is present in the COGRAM rule set (Adriaens and Macken, 1995: 130). This rule was evaluated with four examples, and returned a French generic score that was superior to the German generic score (0.625 compared to 0.125). Differences between the two languages were also confirmed by two 'extra' scores of level 2 obtained in French, suggesting that only marginal improvements were introduced by the rule. This may be explained by the fact that in the four examples the MT system analysed 'since'

126

as a conjunction introducing a subclause of reason in both languages, and not as a temporal subclause, which would have distorted the original meaning of the source input. The superior generic score for French may be explained by the more idiomatic translation of 'because' by 'parce que' instead of 'puisque' for 'since'.

The effectiveness of rule 33, *'Use a comma before a coordinated clause'*, was evaluated using six examples. Only one of them returned a positive 'extra' score (example 196 for the English-French language pair):

⊠ Enter addresses one at a time and separate them by a space.

A. Saisissez les adresses une par une et séparées elles par un espace.

☑ Enter addresses one at a time, and separate them by a space.

B Saisissez les adresses une par une, et séparez-les par un espace.

In the above example, the MT system incorrectly parsed the pre-CL segment, since 'separate' was translated as an adjective. The introduction of the comma in the post-CL segment removed the ambiguity. In this particular example, the second verb of the pre-CL segment was also mistranslated in German MT output A. The effectiveness of this rule may not be frequent in both language pairs, but can greatly improve the comprehensibility of MT output in certain cases.

The effectiveness of rule 21, *'Do not use the slash to list lexical items'*, was evaluated with seven examples. Five of them returned positive 'extra' scores in French, whereas only one of them returned a positive 'extra' score in German. The score differences may be explained by different analysis and transfer rules in the two language pairs. For instance, MT outputs A differ greatly for example 124:

⊠ Enhancements/Problems fixed

A. Les perfectionnements/problèmes les ont résolu

A. Verbesserungen/Probleme behoben

Whereas the German word order follows the English word order, extra words have been introduced in the French output, which is not comprehensible. In this particular example, the rule is more effective for French than German because German MT output A was much more comprehensible than French MT output A. One example, however, returned positive 'extra' scores in both languages (example

127). As a conclusion, the positive impact of this rule can be equally important in both languages, especially when the slash is used to separate two nouns or an adjective and a noun. This is consistent with O'Brien's conclusions (2006: 170), who also found that improvements are limited when the slash separates other lexical items, such as two conjunctions (and/or).

Rule 11, '*When appropriate, use an article (a, an, the) before a noun*', returned 3 positive 'extra' scores in French whereas it did not return any in German. As stated by Heald and Zajac (1996: 207), 'the explicit use of the article is motivated by the assumption that it makes understanding easier and helps to clear possible confusion between noun and verbs.' This assumption was confirmed by the positive 'extra' scores of level 1 obtained for examples 49 and 56 in French. In these two examples, nouns requiring an article followed ambiguous words: an –ing word in example 49 and a subjectless third person verb in the present tense in example 56:

☒ The problem might be a result of upgrading your operating system without first uninstalling programs that are not compatible with Windows XP.

☒ Blocks repeated Internet attacks.

For example 56, the rule was more effective in French than German because the ambiguous verb was mistranslated in French MT output A, whereas it was properly translated in German MT output A. In example 49, the ambiguous –ing word was mistranslated in both languages, but the complexity of the sentence masked the improvement in German. In short, results showed that the effectiveness of this rule is not consistent, despite its potential positive impact on the comprehensibility of MT output. It should be noted, however, that the description of this rule appears too restrictive. In certain cases, such as example 56, using the article 'the' in the post-CL segment is incorrect since the noun phrase contained in the segment has not been explicitly defined. A determiner such as 'any' would be more natural in the source text but may cause transfer problems due its various translations into both French and German. This example suggests that articles such as 'the' have the potential to remove ambiguity, but sometimes at a cost. Since the acceptability of the source text seems to be impacted in certain cases, considering an automatic insertion of articles as part of a pre-processing process may be envisaged. Such a process, which has been suggested by Nyberg et al. (2003: 276), requires further studies.

The effectiveness of rule 13, *'Do not use the impersonal passive voice'* was evaluated with 3 examples which all contained the same structure, since it was the only structure of this type found in the corpus ('it is recommended'). The rule was not effective in German since the structure contained in the pre-CL segment was correctly handled and translated in German MT outputs A. This was not the case in French MT outputs A, but no positive 'extra' scores were returned despite a high generic score. This may explained by the fact that the alternative structure used in post-CL segment was translated following the word order of the source text, instead of using a more idiomatic formulation ('nous recommendons que' instead of 'nous vous recommendons de'). Based on the examples used in this part study, the impact of this rule is limited to minor improvements that do not guarantee the comprehensibility of the French MT output.

The effectiveness of rule 16, *'Avoid the use of progressive verb participles'* was evaluated using nine examples. Eight of them contained a present progressive verb, and one of them contained a past progressive verb. The only example to return a positive 'extra' score in French was the example containing a past progressive verb (which is unusual for this type of verb in English). The translation of the pre-CL segment resulted in a French imperfect tense, which was not as accurate as the passé composé produced in MT output B. This minor improvement must be contrasted with the deterioration noted with one of the rewritings in German (example 92). The reduction of the unambiguous progressive form from the pre-CL segment into a simple form introduced ambiguity in the post-CL segment. The effectiveness of this rule is therefore extremely limited and could have negative impact in certain cases.

Differences in the handling of specific structures across language pairs was noted with results obtained for rule 31, *'Whenever possible, the use of "wh-" questions should be avoided'*. Such a rule does not seem to address any ambiguity, but rather to address potential deficiencies of MT systems. This rule proved much more effective for French than for German, because German MT outputs A were much more comprehensible than French MT outputs A. Median scores for German MT outputs A were {3.5, 4, 4, 3} compared to {2.5, 2.5, 2.5, 1.5}. These results show that simple and unambiguous interrogative structures were not handled properly in

the English-French language pair. The rewritings introduced in post-CL segments therefore improved the comprehensibility of French MT output.

The effectiveness of rule 50, *'Keep both parts of a verb together'* was evaluated with four examples. It returned two 'extra' scores in French but only one in German. A negative 'extra' score was also returned in German for example 272 in which two conjoined particles were used in the pre-CL segment. The results obtained for this last example suggest that the use of conjoined particles is handled more easily in German than in French, due to the fact that particles can also be used in this language. Overall, this rule may be effective in both languages, but its application may conflict with the application of rule 52, *'Use single-word verbs'*.

Rule 26, *'Avoid the ellipsis of verb'*, was evaluated with six examples, due to the various contexts in which verbal ellipses can occur, whether in a single sentence (as in example 155), or across two sentences (as in example 154). Verbal ellipses can be found in patterns such as 'if it does not,' or 'but the files are not,' in which the lexical verb is omitted. More 'extra' scores were obtained in French. All six examples were poorly handled in MT output A in French and German, except for 'if not', which returned an idiomatic translation in French, 'sinon' in example 151. This example was a special case, because both the verb and the subject were omitted. Ellipses of subject are discussed next.

Rule 27, *'Avoid the ellipsis of subject'*, covered two distinct structures: subject ellipsis and a combination of subject and 'verbal operator ellipsis'. The latter structure is classified by Halliday and Hasan (1976: 174) as a type of verbal ellipsis. In this structure, however, the subject is also omitted from the clause, so it was decided to include it in rule 27 instead of rule 26. This second structure differs from the first one because both the subject and the operator have to be re-inserted in the reformulation, which then contains a finite dependent clause. Segments 156, 159, 165, and 166 were selected to cover the first type of structure, while the other 10 segments were chosen to evaluate the impact of operator ellipsis on the comprehensibility of MT output. This rule returned more 'extra' scores for French than for German (6 compared to 1). The generic scores were also in favour of French (0.8 for French and 0.1 for German). The results of this rule, however,

should be analysed in the light of the results obtained for rule 36. When testing rule 36, it was noted that pronouns were automatically inserted in German output. The same mechanism was again used by the MT system to translate some of the pre-CL segments in German. The mechanism worked well when the subject was the same in both clauses. For instance, examples 163 and 164 returned identical German MT outputs regardless of the type of source segment (pre-CL or post-CL). This is explained by the correct insertion of subjects in the German dependent clauses. The same insertion is not performed in French, so the comprehensibility of MT outputs A was greatly impacted, as shown with example 164:

- ☒ Click OK when asked to confirm the deletion.
- A Cliquez sur OK une fois demandé à confirmer la suppression.
- ☑ Click OK when you are asked to confirm the deletion.
- B Cliquez sur OK quand vous êtes invité à confirmer la suppression.

The above example confirms that the comprehensibility of French MT output can be greatly improved by the introduction of rule 27. This example also highlights the differences that exist between the MT system's transfer rules. However, the automatic insertion of pronouns in German output does not work when the source input is ambiguous, as shown in example 168:

- ☒ Verify that the item is checked. If not checked, then check it.
- A Überprüfen Sie, dass das Element überprüft wird. Wenn Sie nicht überprüft werden, überprüfen Sie es dann.
- ☑ Verify that the item is checked. If the item is not checked, then check it.
- B Überprüfen Sie, dass das Element überprüft wird. Wenn das Element nicht überprüft wird, dann überprüfen Sie es.

In the above example, the pronoun 'Sie' was automatically inserted in MT output A, probably due to the presence of the imperative verb 'check' in the main clause. In this example, however, the ellipsis concerned a noun from the previous sentence: 'item'. This suggests that ellipses should not span over several sentences, despite the fact that the effectiveness of the rule was completely masked by the mistranslation of the verb 'check'.

Rule 47, *'Move document names, error messages, or section titles to independent segments'* is based on the textual feature of technical support documentation identified in section 1.3.4 of Chapter 1. This rule was evaluated with six examples, and returned a high number of positive 'extra' scores for French (4). Results were not as clear-cut for German due to the inherent complexity of certain examples (such as 257 or 260). Besides, example 259 revealed that short embedded titles marked with distinct punctuation marks (such as pair of double quotes) were not as problematic for German as they were for French. The comprehensibility of the MT output in this particular example was improved with the introducion of the rule in French, but not in German.

After performing a micro-analysis of all the rules that appeared to be more effective for French than German, it transpires that certain rules can be classified as language-dependent rules. Based on the results obtained with the examples used in the test suite, three rules specifically fall into this category: rule 31, rule 13, and a specific part of rule 16. The rules present in the last category of rules must undergo the same review process to determine whether certain rules improve specifically the English-German language pair.

4.3.6. CL rules that are more effective for German than for French

Based on the generation of generic and 'extra' scores, certain rules appeared to be more effective in German than in French. Rules were initially placed in this category when one of two criteria was met:

- The rule's frequency score for German was greater than or equal to 33 and the rule's frequency score for French was less than 33.

- The rule's frequency score for German was greater than 0 and the rule's frequency score for French was less than or equal to 0.

These rules are listed in Table 4-8 below:

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (-)	French 'extra' score 2 (-)
Rule 3: Do not start a sentence with an implicit subject in a subjectless, non-finite clause premodifying a finite clause, even when the subject is the same in both clauses.	3	20	0	1	0	0	0
Rule 6: Do not omit the subordinating word 'that' after verbs or verbal nouns	3	33	0	1	0	0	0
Rule 8: Do not omit relative pronouns and a form of the verb 'be' in a relative clause. The post-modifier is an -ing word.	5	20	0	1	0	0	0
Rule 15: Do not use an ambiguous form of 'have'.	7	43	0	3	0	0	0
Rule 24: 'Repeat the preposition, conjunction, or infinitive marker in coordinated prepositional phrases, subordinate clauses, or infinitive complements'.	5	10	0	0	1	0	0
Rule 38: Rewrite ing-words that are complements of other verbs.	4	50	0	2	0	0	0
Rule 40: Avoid very short sentences (less than 4 words).	5	20	0	1	0	0	0

Rule number and name	Number of examples	German freq. score	French freq. score	German 'extra' score 1 (+)	German 'extra' score 2 (+)	French 'extra' score 1 (-)	French 'extra' score 2 (-)
Rule 44: Avoid splitting infinitives, unless the emphasis is on the adverb.	5	20	0	1	0	0	0

Table 4-8: Rules with more positive impact in German than in French

Rule 3, *'Do not start a sentence with an implicit subject in a subjectless, non-finite clause premodifying a finite clause, even when the subject is the same in both clauses',* was evaluated with five examples. Only one of them returned a positive 'extra' score in German (example 15). This example was different from the other four, since the subjectless non-finite clause contained two –ing words. While the pre-CL segment was correctly parsed for the English-French language pair, it produced output that received three medium scores in German. In most cases, this rule does not have any impact in both languages. When a subjectless non-finite clause is limited to one –ing word, the comprehensibility of German output seems to be improved.

The effectiveness of rule 6, *'Do not omit the subordinating word "that" after verbs or verbal nouns',* was evaluated using three examples. The rule was not as effective in French as it was in German due to syntactic differences resulting from the absence of the conjunction in the source text. Despite the absence of the conjunction 'que' in the MT output A of example 24, three French evaluators thought the sentence could be understood by end-users. In the other two examples, examples 25 and 26, the subordinating conjunction 'que' was automatically inserted in French output despite being absent from the pre-CL segment. In German, the corresponding conjunction 'dass' was only inserted in example 25. Example 26 therefore returned a positive score of level 1 in German due to significant syntactic differences between MT output A and MT output B. Based on these three examples, the rule appears to be slightly more effective in German than in French.

The effectiveness of rule 8, *'Do not omit relative pronouns and a form of the verb 'be' in a relative clause, when the post-modifier is an -ing word'*, was evaluated with five examples. This rule was not effective for the English-French language pair, and only effective in one example for the English-German language pair. For the English-French language pair, all of the present participles present in pre-CL segments were parsed correctly. For the English-German language pair, full relative clauses were automatically inserted in four of the MT outputs A, despite the absence of relative pronouns in pre-CL segments.

The effectiveness of rule 15, *'Do not use an ambiguous form of 'have'*, was evaluated with seven examples, and returned positive 'extra' scores in German only. The negative impact originating from the ambiguities that the rule tries to address was more visible in German MT outputs A than in French MT outputs A. In the pre-CL segments of examples 82, 83, 84 and 85, the structure 'have' + past participle is ambiguous, because it could be analysed as a causative structure in which the past-participle is used in a passive way. This is shown with the pre-CL segment of example 82:

☒ Symantec does not recommend having multiple firewalls installed on the same computer.

This example returned a positive 'extra' score of level 1 in German, suggesting that a rewriting using the active voice could improve the comprehensibility of the MT output. A similar positive 'extra' score was obtained with example 86, which did contain a causative structure. This structure failed to be transferred accurately in German, which impacted on the comprehensibility of MT output A. The results obtained with these seven examples suggest that the rule was more effective for German than French. When examining other segments present in the test suite, however, another ambiguous form of 'have' was found in one of the post-CL segments (in example 153). This segment contained a negative form of this ambiguous structure, which was mistranslated in French output:

☒ If you do not have Norton SystemWorks installed, then contact your computer manufacturer.

A Si vous ne faites pas installer Norton SystemWorks, contactez alors votre fabriquant informatique.

This example suggests that the rule could also be effective for the English-French language pair when negative structures are used.

The effectiveness of rule 24, *'Repeat the preposition, conjunction, or infinitive marker in coordinated prepositional phrases, subordinate clauses, or infinitive complements'*, was evaluated using five examples. Only one of them returned a positive 'extra' score (of level 2) in German. The pre-CL segment of example 144, which contained conjoined prepositional phrases separated by a comma, is presented below with the two MT outputs A:

☒ The following solutions have worked for some of our customers, but not all of them.

A Die folgenden Lösungen haben für einige unserer Kunden, aber nicht alle funktioniert.

A Les solutions suivantes ont fonctionné pour certains de nos clients, mais pour pas tous.

While three German evaluators attributed 'Good' scores to MT output A, three French evaluators gave 'Excellent' scores despite a word order problem ('pour pas') introduced by the absence of the preposition in the pre-CL segment. Based on this analysis, it cannot be stated that the rule is more efficient for German than for French. The effectiveness of this rule, however, was not consistent across the three examples containing conjoined phrases that were not separated with commas. In one of the examples (example 143), the preposition was introduced automatically in MT output A despite being absent from the pre-CL segment. To maximise the effectiveness of this rule, its scope could be reduced to conjoined phrases separated with commas.

The effectiveness of rule 40, *'Avoid very short sentences (less than 4 words)'*, was evaluated using five examples. Only one example returned a positive 'extra' score of level 1 (example 221) in German. In this example, the comprehensibility of German MT output A was affected by the incorrect generation of an infinitive structure (with the repetition of 'zu'), rather than by the length of the pre-CL segment. The scores obtained for the other four examples confirmed the lack of effectiveness of this rule. These four examples contained the word 'click', which had been coded with a high priority in the user dictionary so as to be always recognised as a verb. Without this customisation, results may have been different in short sentences, since 'click' may have been analysed as a noun.

Rule 38, *'Rewrite -ing words that are complements of other verbs'*, is recommended by Bernth and Gdaniec (2001: 184). Depending on the valency of the verb they follow, such -ing words can either introduce a subjectless clause as a direct object, or act as an adjunct in the main clause. This rule was evaluated with 4 examples, and returned two positive 'extra' scores of level 1 in German. One of these examples (example 212) contained a verb whose usual complementation is a subjectless -ing clause, 'continue' (Quirk et al., 1985: 118). This structure, however, was not parsed accurately by the MT system in the pre-CL segment. The comprehensibility of the German translation of the post-CL segment was therefore improved thanks to the rewriting. This improvement was not visible in French, due a lexical ambiguity introduced by the homograph 'test', which was analysed as a noun. This micro-analysis suggests that the rule would have been effective in French with a less ambiguous verb. The handling of the verb 'continue' in example 212 contrasts with the output produced for example 213, which contained a similar verb ('try'). In this example, the structure was properly analysed in both languages. In one of the other examples (example 214), rewriting the -ing word as an infinitive was only possible by changing the verb it preceded ('require' to 'need'). Based on these results, the effectiveness of this rule seems to be limited to specific cases. Besides, one should not underestimate the possible ambiguity that rewritings may introduce. An infinitive clause introduced by 'to' could be analysed as a purposive clause. In short, this rule has limited positive impact in both languages and cannot always be implemented by only rewriting the –ing word.

The effectiveness of rule 44, *'Avoid splitting infinitives, unless the emphasis is on the adverb',* was evaluated using five examples. Only one of them returned a positive 'extra' score of level 1 in German (example 245). The infinitive that was split in the pre-CL segment of this example was incorrectly translated as a noun in German MT output A. This was not the case in French, so the effectiveness of this rule was not visible for the English-French language pair in this particular example. In the four other examples, the rewriting did not improve the comprehensibility of the German output because German MT outputs A and MT outputs B were identical regardless of the position of the adverb in the source segment. Based on these results alone, it is not possible to conclude that this rule would always be more efficient for German than for French.

4.4. Findings

The main conclusion that can be derived from the first part of this study is that the effectiveness of individual CL rules is not homogeneous. Not one single rule was always consistently effective in both language pairs. This finding may be explained by several reasons that will be discussed in sections 4.4.1 to 4.4.8:

- The improvement introduced by the CL rule is masked by translation problems triggered by other linguistic phenomena present in the source segment (complexity, ambiguity, homography).

- The improvement introduced by the CL rule is masked or blurred by translation problems triggered by uncontrollable phenomena (morphological problems, duplicated words, wrong personal pronouns)

- The improvement introduced by the CL rule is masked by scoring inconsistencies

- The rule is not effective with specific language pairs because the MT system properly handled the linguistic phenomenon that the rule tried to address

- The rule is only effective with certain examples because the scope of the rule is too large for the language pairs and MT system selected in this study

- The effectiveness of a rule cannot always be reproduced because a standard reformulation is sometimes not available

- The impact of a CL rule may be negative with certain examples if ambiguity is introduced in the reformulation

- The rule is only effective in one language pair because analysis and transfer rules differ from one language pair to the next

4.4.1. Linguistic phenomena masking the effectiveness of an individual rule

Despite attempting to remove certain complex and ambiguous structures from pre-CL and post-CL segments, certain linguistic phenomena masked the effectiveness of certain rules. Leaving certain problematic linguistic phenomena in the source text was, however, sometimes necessary to avoid undermining the external validity of the study. If only simple sentences had been used in the test suite, it would be difficult to generalise and state that the rule is effective for a particular text type.

For instance, a problematic instance of polysemy was noted with example 168, which contained two verbs, 'check' and 'verify', that resulted in the same translation in German, 'überprüfen'. However, one of the source verbs, 'check', was used in the sense of 'activate'. The effectiveness of rule 26 may have been more visible with this example if a different translation had been entered in the German user dictionary.

A problematic instance of homography was also found with the word 'copy', which was sometimes analysed as noun instead of a verb (in example 203 for French) or as verb instead of a noun (in example 202 for German). These lexical problems, which can be exacerbated by a lack of terminology standardisation, were not addressed within the scope of this study. They confirm, however, the assumption that was made before embarking on the design of the test suite: the use of standardised source and target terms would improve the comprehensibility of MT output. This recommendation will be retained for the next part of the study.

4.4.2. Extra phenomena affecting the effectiveness of an individual rule

Besides linguistic 'noise' introduced by phenomena such as complexity and ambiguity (be it lexical, syntactic, or referential), certain translation problems originated unexpectedly in the MT output. For instance, in example 215, the personal pronoun 'you' was translated in the singular form in French ('te') despite the researcher having selected specific translation options prior to the translation process. In other places, certain words were sometimes duplicated for no apparent reason, such as 'zu' in example 221. The duplication of such a word in German MT output A impacted the comprehensibility of MT output A, and may have suggested erroneously that the rule was effective in this particular example. These problems were different from the word order problem mentioned in section 3.3.1.2.1, so they were not fixed prior to the evaluation process.

4.4.3. Ambiguity correctly handled by the MT system

As mentioned during the review of the evaluation results, certain rules were sometimes ineffective because the MT system correctly disambiguated the pre-CL segment and produced comprehensible translations. For instance, this was the case with example 56 in German, where the homograph 'blocks' was correctly analysed as a third person verb instead of a plural noun. This was also the case with most of the examples used to evaluate the effectiveness of rule 39, *'Always write "in order to" before an infinitive in a purpose clause instead of just "to"'*. Other examples found in the test suite, however, showed that the disambiguation was not always performed successfully (examples 103 and 179). Based on this discovery, it seems that the rule would sometimes be effective. Deciding whether it is worth implementing such a rule despite its infrequent effectiveness will then vary from one environment to the next.

4.4.4. Scoring inconsistencies among evaluators

Despite all of the efforts that were made to ensure consistent and reliable evaluation results, certain scores provided by the evaluators sometimes did not correspond to the evaluation criteria that had been defined. Individual scoring inconsistencies were avoided in the analysis of the results by focusing specifically on the examples for which a majority of evaluators agreed. This approach, however, was sometimes

not sufficient to eliminate scoring inconsistencies. For instance, this was the case with the French MT output A of example 144, which received three 'Excellent' scores from French evaluators despite being syntactically incorrect. This was also the case with the French MT output A of example 1, which received two 'Excellent' scores despite failing to convey information accurately. The improvements brought about by the introduction of the rule were therefore not visible when calculating positive 'extra' scores for these particular examples. It should be said, however, that such problematic scoring inconsistencies involving a majority of evaluators were rare.

4.4.5. Scope of the initial rule

Certain rules were only effective with specific examples because the scope of the initial rule was too wide to be effective consistently with the MT system selected in this study. This was the case with rule 16, which only proved effective for the English-French language pair when the example contained a verb in a past progressive form. This was also the case with rule 3, *'Do not start a sentence with an implicit subject in a subjectless, non-finite clause premodifying a finite clause, even when the subject is the same in both clauses'*. This rule was only effective for the English-German language pair when the example contained two subjectless non-finite clauses at the start of a sentence. This issue was also noted when evaluating the effectiveness of rule 36, *'The subject of a non-finite clause must be the same as that of the main clause'*. Final conclusions for this particular rule varied based on the type of non-finite clause present in the examples. Machine-oriented rules must therefore be defined more strictly if their effectiveness is to be evaluated more consistently.

4.4.6. Range of reformulations

Closely related to the issues discussed in section 4.4.1 and 4.4.3 is the issue associated with multiple rewriting alternatives. The range of possible reformulations associated with certain rules sometimes prevented the effectiveness of a rule from being visible with all of the examples selected. For instance, this was the case with rule 5, *'Do not use more than 25 words per sentence'*. Determining precisely whether shorter sentences will always improve the comprehensibility of MT output seems difficult to achieve due to the various ways in which long sentences can be

reformulated. Finding a correlation between sentence length and MT output quality for individual sentences is a challenge that had been identified by Gdaniec (1994: 100). Stricter rewrite recommendations are therefore necessary to improve the consistency of reformulations.

4.4.7. The impact of a rule may be negative if ambiguity is introduced in the reformulation

Some of the results obtained in this study also showed that certain reformulations could be more ambiguous than the original segments. The comprehensibility of MT output B was sometimes negatively affected by the introduction of a rule. This was visible with some of the examples used to evaluate the effectiveness of rule 12, *'Use the active voice when you know who or what did the action'*. Since the effectiveness of this rule had to be evaluated individually, no extra words (such as articles) were introduced in the post-CL segment. Whereas nouns and verbs were correctly disambiguated in the passive structure contained in certain pre-CL segments, ambiguity was introduced in certain post-CL segments with the use of the active voice. In order to avoid such side-effects, the scope of the rule could be reduced or be used in conjunction with rule 11.

4.4.8. The effectiveness of the rule is visible in one language pair only

Based on the results obtained in this study, certain rules clearly improved the comprehensibility of the output corresponding to a specific language pair. This was the case with rule 31, *'Whenever possible, the use of "wh-" questions should be avoided'* when its effectiveness was evaluated with the English-French language pair. However, such a rule was not effective with the English-German language pair because the pre-CL segments were correctly handled by the MT system. A similar phenomenon was also noted with some of the examples used to evaluate rule 6, *'Do not omit the subordinating word 'that' after verbs or verbal nouns'*. The translation of the missing subordinating word was not automatically introduced as frequently in German MT output A as in French MT output A. The rule was therefore more effective for the English-German language pair. The maturity of a language pair should therefore not be underestimated in the evaluation of the effectiveness of CL

rules. Evaluations with more language pairs are required to determine whether the results obtained in this study would apply to different language pairs.

4.4.9. Selecting a final list of rules

Despite the challenges encountered during the evaluation of individual rules and summarised in the previous sections (4.4.1 to 4.4.8), a final distribution of rules can be drawn based on their impact on the comprehensibility of French and German MT output. The 28 most effective rules are presented in Table 4-9 (in no particular order).

CL rule name
Rule 1: Avoid ambiguous coordinations by repeating the head noun, or by changing the word order.
Rule 2: Do not use ambiguous attachments of non-finite clauses (present-participles).
Rule 4: Do not omit hyphens in Noun + Adjective or Noun + Past Participle structures
Rule 5: Do not use more than 25 words per sentence as long as the short sentence is not ambiguous.
Rule 6: Do not omit the subordinating word 'that' after verbs or verbal nouns.
Rule 9: Do not omit relative pronouns and a form of the verb 'be' in a relative clause. The post-modifier is a past-participle.
Rule 14: Do not make noun clusters of more than three nouns.
Rule 15: Do not use an ambiguous form of 'have'.
Rule 17: Do not use a pronoun when the pronoun does not refer to the noun it immediately follows.
Rule 19: Do not use pronouns that have no specific referent.
Rule 20: Do not use an ambiguous form of 'could'.
Rule 21: Do not use the slash to list lexical items.
Rule 22: Do not coordinate verbs or verbal phrases when the verbs do not have the same valency.
Rule 25: Two parts of a conjoined sentence should be of the same type.
Rule 26: Avoid the ellipsis of verb.
Rule 27: Avoid the ellipsis of subject.
Rule 28: Omit unnecessary words as long as ambiguity is not introduced.
Rule 29: Place a purpose clause before the main clause.
Rule 35: Avoid unusual punctuation (dashes with no surrounding spaces).
Rule 36: The subject of a non-finite clause must be the same as that of the main clause.
Rule 41: Avoid embedded parenthetical expressions introduced by commas or dashes.
Rule 47: Move document names, error messages, or section titles to independent segments.
Rule 48: Use a question mark only at the end of a direct question.
Rule 50: If a single-word verb cannot be used, keep both parts of a verb together.
Rule 51: Avoid –ing words in specific situations (gerunds as subjects, coordinated –ing words).
Rule 52: Use single-word verbs.
Rule 53: Avoid ungrammatical constructions (agreement and possessive).
Rule 54: Check the spelling.

Table 4-9: Final selection of CL rules based on their effectiveness

143

The final selection of rules presented in Table 4-9 is based on an evaluation that was performed using a specific MT system and specific language pairs. These rules were selected based on the following criteria:

- the rule has to return at least one 'extra' score in both language pairs (an 'extra' score of category 1 or 2). One exception to this criterion concerns rule 6 *'Do not omit the subordinating word "that" after verbs or verbal nouns',* since the micro-evaluation of example 24 revealed that the well-formedness of the French MT output could be improved with the application of this rule.

- the rule does not have any negative impact in either target language. One exception to this criterion concern rule 15, *'Do not use an ambiguous form of "have"',* which returned a high number of 'extra' scores in German, and showed some potential improvement in French during the micro-evaluation of example 153.

- the rule does not conflict with a rule that has a major impact for both languages (for this reason, rule 18, *'Do not use pronouns',* was discarded since it conflicts with reformulations triggered by rule 22, *' Do not coordinate verbs or verbal phrases when the verbs do not have the same valency'.*

Before further claims can be made on the effectiveness of these rules, their effectiveness must be evaluated at the document level. Some of these rules will be used in the second part of this study if these rules are violated in the documents used for the online experiment. This strategy will helps us determine whether their application can improve the usefulness, comprehensibility, and acceptability of MT documents.

4.5. Summary

In this chapter, an evaluation of the effectiveness of CL rules on MT output was performed, leading to four main findings. Firstly, a limited number of CL rules proved frequently effective for the two language pairs used in this study. Secondly, some of the rules need to accommodate exceptions. For instance, the LSP used in the corpus could not operate without key verbs such as 'back up' or 'log on', despite conflicting with one of the most effective rules (rule 52). Thirdly, certain CL rules, such as rules 7, 8, and 9 have traditionally been considered as guarantees to improve machine translatability. The findings, however, showed that the scope of such rules could be restricted to ambiguous cases only, to avoid a situation where the style of the source text is negatively affected without necessarily improving the MT output. Finally, only a few rules improved exclusively MT output in one target language. These rules can still be applied to the source text as long as they have no negative impact on the other language pair.

Chapter 5

Chapter 5: Setting up an online experiment to evaluate the impact of CL rules at the document level

5.1. Objectives of the present chapter

At the end of Chapter 4, effective CL rules were classified based on the type of impact they had on the comprehensibility of MT output at the segment level. In the present chapter, a number of technical support documents are extracted from the corpus, so as to evaluate the impact of some of these rules at the document level. The methodology used to achieve this objective is discussed in section 5.2. An online experiment using a customer satisfaction questionnaire is conducted to collect users' reactions to documents containing machine-translated content. The first part of this chapter will focus on the methodological decisions made during the design of the online experiment. After reviewing related studies to optimise the methodology used in this study, the final experiment setup is described in greater detail, along with the strategy employed to machine-translate the documents that will be presented to genuine users. The second part of this chapter is concerned with an examination of the documents selected for the experiment. This examination is performed to determine the terminological entries that should be encoded in the MT system's user dictionaries, and identify rule violations present in the original documents. Reformulations will then be applied to original documents so as to produce a set of controlled documents.

5.2. Methodology selection

5.2.1. Review of data collection techniques

As noted in the introduction, no empirical study has ever been specifically conducted to evaluate the effectiveness of CL rules on the usefulness, comprehensibility, and acceptability of machine-translated documents from a Web user's perspective. There is therefore no standard methodology to rely on for data collection. In order to evaluate the impact of the independent variable (the presence or absence of CL rule violations) on the three dependent variables examined, several related studies are reviewed to determine the most suited methodological approach. A standard methodology may not be available, but certain components

may be reused from previous related studies. Two related studies, which focus on the use of SE in technical documentation (Shubert et al.: 1995; Spyridakis et al.: 1997), are first reviewed. Another related study is then reviewed in the field of accessibility and acceptability in technical documentation (Lassen, 2003). Finally, two related studies are reviewed in the field of online machine-translated document publication and customer satisfaction (Richardson, 2004; Jaeger, 2004).

5.2.1.1. Review of studies focusing on source documents

As discussed in section 2.3.1.1 during a review of CL evaluation methods, the effectiveness of a set of CL rules on the comprehensibility of a document has sometimes been evaluated using comprehension tests (Shubert et al., 1995). The main strength of this experiment was its use of genuine SE and non-SE documents - two documents were used, each one containing two versions of the same procedure. The impact of the SE rule set was then evaluated by comparing comprehension results obtained from various groups of subjects. In this study, the focus was placed on the control of the subjects since 121 undergraduate students were used instead of genuine users. This approach did not follow the recommendation made by King (1996: 198), who remarks that an evaluation study based on a comprehension test should involve 'consumers as long as they can be identified'. Overall, the type of experiment used by Shubert et al. (ibid) seems to be a good starting point for the present study as long as subjects are genuine users of technical documentation. Using two different versions of the same document and comparing the reactions obtained from two groups of subjects provides a sound basis to determine if the two groups receive MT documents differently. Using two documents also seems advantageous to ensure that results can be reproduced from one document to the next. A comprehension test, however, does not appear appropriate to obtain reactions from users, as Holmback et al. (1996: 176) state that 'taking a comprehension test is probably not the best way to measure comprehension of a procedure (...) since procedures are written to be performed, not quizzed'. This remark suggests that a usability study could be performed to determine whether the application of CL rules in procedural text can have an impact on the execution of this procedure (such as the speed with which it is executed). However, a laboratory-based usability study presents disadvantages, such as the unfamiliar setting in which users will be monitored and the small samples of subjects used (Spyridakis et al.,

2005: 246). As mentioned in the introduction, a usability study is not appropriate within the scope of this study. It is, however, essential that the evaluation of documents is performed by genuine users of such documents to maximise the ecological validity of the study. Genuine users consult online technical support documents because they have encountered a problem or because they have a question that remains unanswered. Genuine users are therefore different from existing customers who could be contacted and asked to evaluate a document that does not concern them.

For instance, this was the approach that was taken by Lassen (2003: 79) who used an offline survey to investigate user attitudes to the notions of accessibility and acceptability in technical documentation. In order to minimise non-response rates, she contacted prospective respondents before mailing them a questionnaire. Despite using this approach, response rates were around 14% (ibid: 80). Using an offline approach presents certain advantages, especially with regard to time constraints. Gillham (2000: 7) mentions that offline surveys generate "less pressure for an immediate response since respondents can answer in their own time". To some extent, such a setup could be reproduced in an online environment, whereby experimental machine-translated documents could be posted online. Fake users could then be contacted and asked to give their opinions on certain aspects of the translated content. With such an approach, however, the real life scenario discussed earlier would not be in place. If genuine users are to be used within the context of an experimental design, the data collection instrument must be part of the material that is being evaluated. In an online context, this technique is often implemented with customer satisfaction feedback mechanisms. Two examples of studies that made use of customer satisfaction rates to evaluate certain characteristics of machine-translated content are reviewed in the next section.

5.2.1.2. Review of studies focusing on translated documents

Richardson (2004) reports on a project conducted at Microsoft in collaboration with the group responsible for the translation of Knowledge Base product support articles. In this project, more than a hundred thousand product support articles were translated into Spanish and Japanese using Microsoft's MT system, MSR-MT (Richardson et al., 2001). During a four-month pilot study, Robinson (2004: 247) mentions that feedback was obtained from Spanish users by 'surveying a small sample of the approximately 60,000 visits to the web site'. During this period, 380 surveys were obtained, but it is not clear whether all visitors were given the chance to fill in the survey. If they were, the response rate is 0.65% for Spanish and 1.5% for Japanese, suggesting that users' feedback is difficult to obtain. These Microsoft users were asked to evaluate certain characteristics of the machine-translated Knowledge Base content. Richardson does not provide the exact questions that users were asked to answer, but provides rates in the following categories (Richardson, 2004: 248):

- % of customers who are satisfied with KB

- % of customers who were helped to solve their issues using KB (usefulness rate)

- % of customers who thought information is easy to understand

These categories are interesting because they share some similarities with the three variables that are of interest in this study (acceptability, usefulness, and comprehensibility). One of these variables, usefulness, is also found in a study that was conducted at Cisco (Jaeger, 2004). In October 2003, Cisco used a customised Systran MT system to translate more than 4,500 technical support documents into Japanese. Once the online translations were published, a Web survey was used to determine the degree of usefulness of these MT documents for Japanese customers. In January 2004, Jaeger reports that 461 responses were obtained. Again, it is not clear whether all of the approximately 100,000 visitors were asked to fill in a customer satisfaction survey. If they were, the response rate for this particular study is slightly less than 0.5%.

In both the Microsoft study and the Cisco study, the questions used in customer satisfaction questionnaires seem to have focused on the characteristics of the whole knowledge base's content rather than on specific documents. In the present study, questions must relate to specific documents. The experiment's design and instrument are discussed in turn in the next two sections.

5.2.2. Design of an online experiment

Based on the review of related studies, a combination of several approaches has been identified to achieve the second objective of the study. An online experiment must be set up, whereby users will provide their reactions to machine-translated documents by filling in a questionnaire. The main advantages of conducting an online experiment (or Web experiment) have been highlighted by Reips (2002: 245):

- ease of access to a large number of diverse participants

- ease of access for participants with voluntary participation

- high external validity

- avoidance of time constraints

Whereas a laboratory-based experiment would be limited to a small number of participants evolving in an artificial environment, an online experiment can reach many participants with diverse backgrounds and technical settings. The use of this approach guarantees the ecological validity of the experiment and ensures that the findings can be generalised to the population.

Hewson et al. (2003: 48) state that 'in an experimental design the researcher manipulates the independent variable(s) in order to measure the effect on the dependent variables'. In the present study, the feedback provided by users will be analysed to determine whether the introduction of CL rules (the independent variable) has any effect on the following three dependent variables: usefulness, comprehensibility, and acceptability of the machine-translated documents. The independent variable will be manipulated by using certain documents that contain CL rule violations and others that do not.

One possible way to set up the experiment would be to select a specific number of documents from the corpus, and apply certain CL rules to half of these documents. The documents would then be published, and users asked to give their reactions to these documents. Differences between the reactions obtained for controlled documents and original documents would then be measured. This approach would, however, undermine the internal validity of the study, because of confounding variables. It would be extremely difficult to draw conclusions with regard to the effectiveness of the rules if both sets of documents were not equal before the application of the rules.

The approach used in this study is to publish the same number of original and controlled versions of a set of documents. When users access these documents, they will then be randomly provided with one of the two versions and asked to fill in a questionnaire based on their experience. Since this evaluation exercise should be as realistic as possible, they will not be asked to compare both versions. Their experience will be as close as possible to what it would have been with human translation. The only difference will consist in the presence of a disclaimer at the start of the document to warn them that they are about to read a machine-translated document. This disclaimer will also insist on the significance of their feedback, and on the fact that the machine-translated version of the document is provided within the framework of a pilot project. According to Reid (2000: 160), these two requirements are among the 'basic rules in persuading customers to provide feedback'. Customers must be told that their opinion is valued, and must know why the information is required (ibid). Adhering to these principles is essential to ensure that both parties fully understand the implications of such an evaluation.

Figure 5-1 presents the disclaimer that is used in the French documents:
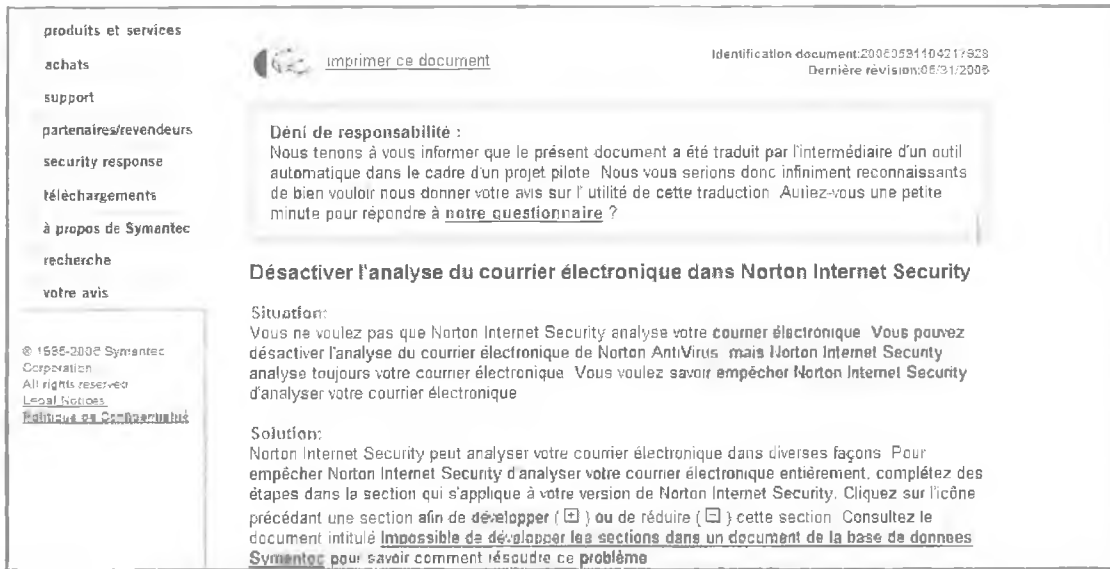
Figure 5-1: French disclaimer used within a custom template

Figure 5-1 shows that users will know prior to their consulting of the document that their feedback is required. This is necessary to maximise the number of responses submitted. Obtaining significant response rates is one of the main objectives of survey-based research. One of the issues with questionnaires mentioned by Gillham (2000: 10), however, is the potential lack of motivation from respondents. According to him, "few people are strongly motivated by questionnaires unless they can see it as having personal relevance". In order to encourage users to provide feedback, it was envisaged to use an incentive by giving away a prize. A study (Göritz, 2006) found that material incentives increase response and decrease dropout in Web surveys. A draw would then have been based on the collection of all respondents' answers. This meant that customers' identities had to be collected, which may have discouraged some customers. If only email addresses had been collected, certain users could have been tempted to submit several forms using different email addresses to maximise their chances of success. In the end, it was decided not to use an incentive, but to emphasise the altruistic appeal. Section 5.2.3 covers in detail the instrument that will be used to collect feedback from users.

5.2.3. Instrumentation

5.2.3.1. Pros and cons of customer satisfaction feedback mechanisms

Symantec's technical support group uses a Web-based feedback mechanism that allows them to monitor their customers' satisfaction levels, as shown by Figure 5-2.

Figure 5-2: Customer satisfaction form for technical support documents

This feedback mechanism uses a combination of closed and open questions, which make it difficult to obtain specific answers. The question 'Was your question answered by this document?' seems too precise because it allows respondents to submit answers without indicating whether the document was useful. A document may not answer a specific question, but it may be useful by pointing to other reference materials. On the other hand, the free text satisfaction box does not encourage precise answers. For desperate or frustrated customers, this may even be seen as the only way to request help from the organisation. This was confirmed by examining some of the feedback received in previous months. Valid remarks were mixed with a large number of technical support requests, such as 'please call me at this number to tell me how to...'.

Such a feedback mechanism is in place because 'it is generally agreed that the organization must listen to customers to improve the quality of goods and services.' (Armistead & Clark, 1992: 1). Of course, the feedback is provided by customers on a voluntary basis, and real constructive feedback may be scarce. Since there is no special incentive for users to provide feedback – apart from a promise of improved services – satisfied customers may not feel the urge to state that the document they consulted answered their questions. In most cases, their online exposure to the technical support document was triggered by a problem they did not expect in the first place, so why should they feel grateful to the company from which they bought the product? On the other hand, dissatisfied customers can leave the Web page without making a comment. This makes it difficult to understand how the service could be improved. In this study, silence from both satisfied and dissatisfied users could affect the accuracy and reliability of the results. This issue is discussed in the next section.

5.2.3.2. Ensuring the accuracy and reliability of the results

In the present study, the emphasis must be placed on specific characteristics of documents by soliciting feedback from users who will be exposed to such machine-translated documents. This approach presents some issues with regard to the reliability of the results since the quality and veracity of anonymous answers cannot be controlled. However, Hewson et al. (2002: 43) state that 'in general we may assume that our participants are giving genuine and accurate answers' (2003: 44). Besides, using anonymous answers should reduce the problem of social desirability mentioned by De Vaus (2002: 107). Respondents will not feel under any pressure to provide 'good' answers when providing feedback.

Obtaining as many answers as possible is essential to ensure that the statistical power of the experiment is sufficient. Evans et al. (2004:14) define statistical power as 'the ability to detect a difference between treatment groups in an experiment, if a true difference exists'. In order to avoid accepting false hypotheses or rejecting true hypotheses, the statistical power of the given experiment must be sufficient. This can be achieved by increasing the number of respondents. Evans et al. (ibid) state that when the 'power is not high enough, a study may not produce a statistically significant result even when there is a true difference'. Increased accuracy will

therefore be obtained by reducing the sampling error to a minimum. De Vaus (2002: 81) mentions that the sampling error of a sample size of 100 is 10%. The longer the experiment stays online, the more accurate the results, if the number of respondents increases over time. The components of the instrument used in the experiment are described in the next section.

5.2.3.3. Instrumentation design

Using open questions in a Web survey presents disadvantages when there is no control over the selection of respondents. In her study, Galesic (2006: 325) found that 'open questions seem to negatively affect interest in the questions'. In order to maximise response rates, she mentions that respondents must be kept interested. As open questions should be avoided, using a short list of simple closed questions and answers appears to be the most suitable approach. This also follows the recommendation made by Jackob and Zerback (2006: 20). It was decided to use dichotomous responses, whereby respondents would be asked to select one of two alternatives ('Yes' and 'No'). De Vaus (2002: 106) warns that "the danger with using 'don't know' and 'no opinion' alternatives is that some respondents select them out of *laziness*". In order to avoid this problem, only two radio boxes would be displayed along each question. Symantec's existing question ('Did this document answer your question?') was retained and supplemented with four others following a logical flow. These four extra questions were:

- Was this document useful?

- Did the quality of the translation hamper your comprehension?

- Would you have preferred to read the English version of this document?

- If you encountered another problem in the future, would you consult again a machine-translated document from the Symantec Knowledge Base?

The first question is asked to determine whether a machine-translated technical support document can help users solve their problems. Positive answers to this question should show that the use of refined MT output, whether the source document is in its original form or controlled form, goes beyond the mere understanding of the overall theme of the document. It may not be fully automatic

high quality translation, but as Van der Meer (2006) puts it, close to 'fully automatic useful translation'. The difference in terms of 'yes' and 'no' answers should indicate whether the use of CL rules can play a part in the overall cognitive process with which the users have been confronted. Since MT output is not as fluent as human translation, and in some cases, not as accurate, users may find the document useful but at a cost. Due to unusual syntax or word order, it may take them longer to understand the information contained in the document, but still find it useful. The answers to the question 'Did the quality of the translation hamper your comprehension?' should therefore help us to find out whether CL rules can help MT output become more comprehensible, and in turn more acceptable.

As stated in the introduction, the concept of acceptability is tightly intertwined with the concept of usefulness, since the acceptability and use of a product or service are triggered by an 'implicit or explicit cost-benefit analysis' (Shackel, 1990: 32). From this statement, it is inferred that users will find machine-translated documentation acceptable when they tolerate some of the textual disturbances caused by an MT process. In this case, the number of grammatical or stylistic disturbances does not outweigh the number of textual characteristics that users expect to find in technical support documentation. Lassen (2003: 81) states that 'acceptability is an ambiguous notion that may imply grammaticality to some respondents, while it may imply stylistic acceptability to others'. When analysing the results of her survey on the acceptability of technical manuals, Lassen (ibid) found that the term 'acceptable' had not been understood in the same way by all respondents. In order to avoid such a situation in the present study, it was decided to use two separate questions to evaluate the acceptability of machine-translated documentation, without using the term 'acceptable'.

The first of these two questions asks users to state whether they would have preferred to read the English version of the document. This question is being asked because the machine-translated document may lack some of the expected features that make the document a 'specimen of the genre' (Lassen, 2003: 78). Regardless of their level of proficiency in English and despite the overall usefulness of the document, certain users may answer that they prefer reading an English version of this document rather than a machine-translated version in their own language. This

would suggest that the intrinsic quality of a document goes beyond its usefulness. Not only should a document be useful, but it should also include indicators of credibility and stylistic accuracy that meet the expectations of its receivers. This investigation will be reinforced by the final question, which will ask users to state whether they would consult again a machine-translated document in the future. This question is asked to determine how acceptable an MT document is to its receivers since the answers to this question will depend on the results of the users' cost-benefit analysis of the service provided. If MT documents are missing too many key stylistic and grammatical features, it is inferred that they will be rejected by users and not consulted again in the future. The questions used in the French survey are shown in Figure 5-3:

Aidez-nous à mieux connaître vos besoins !

Vous pouvez nous aider à améliorer la qualité de nos prestations en répondant aux questions suivantes :

Ce document vous a-t-il servi ?	◯ Oui	◯ Non
Ce document a-t-il répondu à votre question ?	◯ Oui	◯ Non
La qualité de cette traduction vous-a-t-elle empêché de comprendre ce document ?	◯ Oui	◯ Non
Auriez-vous préféré consulter la version anglaise de ce document ?	◯ Oui	◯ Non
Si vous rencontrez un autre problème à l'avenir, consulteriez-vous de nouveau un document de la base de données de Symantec traduit par un outil automatique ?	◯ Oui	◯ Non

Figure 5-3: French submission form for online survey

Figure 5-3 shows that the design of the questionnaire was kept as simple as possible. This follows the recommendation made by Dillman et al. (1998: 4), who found in a study that response rates could be negatively affected by 'fancy design'. Figure 5-3 also shows that customers are encouraged to provide feedback so that the overall service can be improved in the future. By asking them to rate the usefulness level of MT output, users are implicitly told that the translation coverage of technical support documents might be extended in the future. All documents would be available in English as well as in a range of target languages thanks to MT. Upon submission of this form, users will be of course thanked for the time they have devoted to filling out the questionnaire. Making sure that users know what is going

to be done with the information, and thanking them, are two basic rules of customer feedback generation (Reid, 2000: 160).

Once the overall design of the instrument was completed, it had to be implemented so as to replace the existing feedback form in certain documents taken from the corpus. The next section describes the steps that were taken to put a prototype questionnaire in place, before publishing it and making it available to users.

5.2.4. Technical implementation of the experiment and instrument

As described in Chapter 1, Symantec's technical support documents are generated from various knowledge bases. Documents are stored in language-specific knowledge bases using unique identification numbers and pre-defined templates. These documents may also be referenced from generic technical support Web pages residing on a separate Web server. In order to implement the experiment's specifications outlined in the previous section, customising existing Web pages was not the only task required. One of the objectives was to ensure that users would be provided with either the controlled version or the original version of a machine-translated document. This meant that a randomisation mechanism had to be put in place so that one of the documents would be transparently generated to respond to a user request. To implement this requirement, it was decided that both versions of the document should reside in the same knowledge base location. A piece of JavaScript code would then hide or display one of the two versions depending on a random number generated at request time. Since both documents had to reside in the same location, a custom template had to be created. A prototype containing a controlled section and an original section was crafted and made available for testing on a staging server. The main purpose of this prototype was to ensure that the questionnaire displayed properly in Web browsers, and that answers were sent to the right server.

Using this prototype also proved extremely useful in addressing a number of issues that had not been envisaged during the design phases. Due to the nature of the knowledge base's template, having a controlled section and an original section was not sufficient since both versions of the document happened to be sharing some common components such as the title. This issue was fixed by making sure that

every field of the template would contain two entries for each of the following document's sections:

- Title

- Situation

- Solution

- Technical information

- Survey

Two other issues were identified during the testing phase of the questionnaire prototype. First, it was realised that it would be easy for users to read the document and leave the page without completing the survey. It was therefore decided to use a pop-up mechanism that would trigger if users left the page without completing the survey. The German pop-up is shown in Figure 5.4:

Figure 5-4: German submission form displayed in a pop-up window

Figure 5-4 shows that radio boxes were aligned on the right hand-side of the page to avoid back and forth eye movements. This decision was made based on the study conducted by Bowker and Dillman (2000), who found that left-aligned answers could negatively affect response rates. Of course, displaying this pop-up window does not guarantee that the survey will be completed by all users. Certain Web browsers automatically block such windows, and users have to decide whether they want to have them displayed. Besides, closing such a window is quicker than clicking five times in the radio boxes and once on the 'Submit' button. Yet, users are given the possibility to see that the survey would not take more than thirty seconds of their time to be completed.

The second issue that was noted when the prototype was used is that certain users would only be presented with the original version of the machine-translated document. Should they take the time to complete the survey, they should be rewarded with the possibility to consult the controlled version of the document. It was therefore decided to customise the 'Thank you' message by adding a link to the controlled document. As aforementioned, it was important that users would only provide feedback once, so it was decided to use temporary cookies to prevent them for submitting the form more than once. The possibility of having multiple submissions is sometimes regarded as a disadvantage in a Web experiment (Reips, 2002: 245). Using temporary cookies was selected as a measure to address this problem. Once the questionnaire prototype was fine-tuned, it was embedded within documents selected from the corpus. These documents are provided in Appendix I. The next section describes this selection process in detail.

5.3. Selecting documents for the experimental variations

5.3.1. Criteria for the selection of documents

In the second part of this study, an online experiment must be conducted using at least two documents. Three initial criteria were identified to select these two documents from the corpus for the experiment. First, they must be accessed by a sufficient number of users so that feedback can be obtained. Second, they should ideally not already have been translated into French and German, so that existing human translations do not have to be removed. This solution, however, could be envisaged if the number of answers received during a pilot study suggests that statistical tests may be jeopardised by small cell sizes. Finally, these source documents must violate a sufficient number of CL rules. If only one or two rules are violated, the 'power' of the experiment may not be sufficient enough (Evans et al., 2004: 14).

5.3.1.1. Relevance of the documents for French and German users

In order to determine whether some source documents are likely to be consulted by a large group of users once machine-translated, one can look at the number of Web hits that certain source documents have received from French and German users. Two hypotheses can be drawn from these statistics. First, the people who have accessed a source document in English after reading a document in their own language might have preferred to consult a translated version of this document. They might benefit from a machine-translated document if they failed to fully comprehend the English document. Second, if certain French and German users are looking at specific English documents, one may assume that these documents are not US-centric, and therefore relevant for the French and German localised versions of the product. Providing a translation for these documents might generate some interest from French and German users.

Web hits statistics were obtained for each English source document present in the corpus. These statistics were gathered by querying a tracking database, using the identification number of each corresponding document. As the corpus had been

stored in a static format, updates to original documents had not been monitored. Yet, certain documents had been updated since the start of this study. Before any document could be machine-translated, updates had to be recorded to ensure that all machine-translated content would still be relevant. Other documents had become obsolete or inactive by the time the statistics were harvested, which facilitated the selection process. The documents' identification numbers, however, never changed and provided a reliable method to find out more about the way customers or users accessed a particular document. When accessing a particular English technical document, French or German users could have been in one of the following seven scenarios:

- They were redirected to the document after encountering a problem with their localised program (or agent). An error message asked them to troubleshoot the issue by consulting a particular technical support document. This is the scenario that provides the majority of Web hits.

- They navigated to the document because it was present in a list of 'hot topics' on the French or German support Web site.

- After reading a technical support document in their own language, they decided to read the English version (maybe because the English version was more up-to-date).

- They found a link pointing to the document after querying a knowledge base using the technical support Web site's search engine in their own language.

- They were pointed to the document after reading a document on the Symantec Security Response Web site, which provides information about viruses and vulnerabilities.

- They were redirected to the document after using an Automated Support Assistant, which is a tool used for automatic troubleshooting.

All of the Web hits statistics obtained for these particular sources were added to evaluate the popularity of the corpus' content with French and German customers. This strategy is often used at Symantec to determine whether it is worth translating

a document into a certain language. An analysis of these statistics revealed that few English documents triggered any interest from French and German users. This may be explained by the limited English skills of some of these users, which could prevent them from attempting to consult a document in English. These statistics were considered as an indicative reference only, the real factor being the number of Web hits these documents will receive once they are machine-translated and published in French and German.

Some of the documents receiving a high number of hits, however, were too short and only contained links to other documents. They were not considered for selection because they did not present any comprehensibility challenge. Other documents were already translated into French and German, so they were not initially considered. Some documents were simply too important to be used in such an experiment, such as the document *'Configuring Norton AntiVirus to provide maximum virus protection'*, which deals with a sensitive topic. One should not forget that most users are Symantec customers entitled to technical support. The distinction between users and customers is made by Rice (1997: 6) who writes that 'customers are people who use services and pay for them while users are often people who use the product but do not pay for it'. Since Symantec's technical support Web sites are not restricted to customers through a login access, both customers and users have a chance to get Web-based technical support. By increasing the exposure of content that may not be comprehensible, one faces the risk of infuriating or even losing customers. A bad experience related to an unsolved problem can then be associated with the brand. Since the 'brand image is what distinguishes a product from its competitors' (Rice: 129), risks must be minimised.

The focus was then placed on documents of average size, so that the amount of user dictionary preparation prior to the MT process would be limited to a minimum. Both procedural documents and descriptive documents were considered for selection. However, only procedural documents showed any signs of interest from French and German users. For instance, the descriptive source document, *'Features included in Norton SystemWorks Premier Edition',* did not return any Web hits from French or German users.

In the end, three documents were carefully reviewed based on their number of received Web hits:

- *'Turning on or turning off email scanning in Norton AntiVirus'* received 81 Web hits from French users and 94 from German users over a three-month period[11]

- *'Error: "Norton AntiVirus 2005 has encountered an internal program error." (3009,1003)'* received only 2 Web hits from French users but 107 from German users.

- *'Disabling email scanning in Norton Internet Security'* received 106 Web hits from French users and 107 from German users over a three-month period

The violations of CL rules these documents contained were then examined. After this review, the third document was discarded because it only contained two rule violations. The first document contained 14 rule violations, and the second 9.

5.3.1.2. Conducting a pilot study

Concerns surrounded the selection of the second document based on the low number of received Web hits from French users. It was therefore decided to conduct a pilot study for two weeks to determine whether a sufficient number of responses would be returned to conduct statistical tests. If the number of answers received was bound to jeopardise a statistical analysis, a popular document would be used instead. When analysing the response rates after a two-week period, the lack of feedback was obvious, regardless of the document version presented to users. The second document received 85 Web hits, but only five questionnaires were submitted, as shown in Table 5-1:

Document	Number of hits for the German document	Number of German respondents	Number of hits for the French document	Number of French respondents
NAV05_131	337	12	316	8
Document 2	41	2	44	3

Table 5-1: Number of answers received (May 31st – June 13th 2006)

---

[11] This document is referred to as 'NAV05_131' in the remainder of this dissertation.

Table 5-1 shows low response rates that could jeopardise the statistical analysis that had been planned. In order to circumvent this feedback drought, two decisions were taken. First, it was decided to make the first document more visible on the French and German technical support home pages, as shown by Figure 5-5:
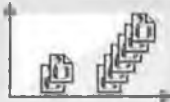

Figure 5-5: Increased visibility of the first document

The second decision was to replace the second document with a more popular document. Based on the low number of responses received for the second document (five in total), it was decided to remove an already-translated popular document, and replace it with machine-translated content.

5.3.1.3. Selecting a popular document

To decide on a popular document for both French and German users, Web hits were used again. Based on the Web traffic monitored between the 6th and the 13th of June, the following document was chosen: "Error: "Norton AntiVirus was unable to scan your Instant Messenger..." (3021,4) after installing Norton AntiVirus" [12]. This document received 1112 visits from French users and 1131 from German users during this timeframe. This high number of visits may be explained by the ease of access to this document and the topic it covers. It contains instructions to help users

---

[12] This document is referred as NAV05_121 in the remainder of this dissertation.

ensure that the information they exchange using a popular instant messaging application (MSN Messenger) is considered safe by the Symantec product. Due to increasing security paranoia, one can easily understand why users would like to solve this problem if they see the error message contained in the title of this document. Provided that they are online - which should be the case since they are using instant messaging technology - clicking a link embedded in the error message will redirect them automatically to the corresponding technical support document. Such high Web hits statistics make it the third most popular document for the French and German Norton AntiVirus knowledge bases, which confirms the risks and the visibility associated with this study.

This section provided some information with regard to the selection and the publishing of documents. The next section will describe the steps that were used to prepare the documents prior to the machine-translation process.

5.3.2. Preparing source texts for MT

How original and controlled document versions should be machine-translated depends on one of the objectives of the present study. As much as possible, this case study is based on a real-life scenario, whereby users are asked to rate a document that has been almost entirely translated by an MT system using a production workflow. Symantec has been using MT for some time, so its current MT process is briefly discussed to determine whether some of its components can be useful for the present study.

5.3.2.1. Leveraging Symantec's MT resources

Symantec's MT process uses a combination of TM and MT technologies whereby a knowledge base document is extracted from its knowledge base in an XML format, and analysed against a Trados TM (Roturier et al., 2005). All source segments that do not return high fuzzy matches above a predefined threshold value are then automatically sent for translation to a Systran 5.0 MT engine. Once the segments are machine-translated, they are automatically imported back into the TM so that translators can translate the new document using a TM that contains a combination of previous matches and machine-translated segments. This process is shown in Figure 5-6:
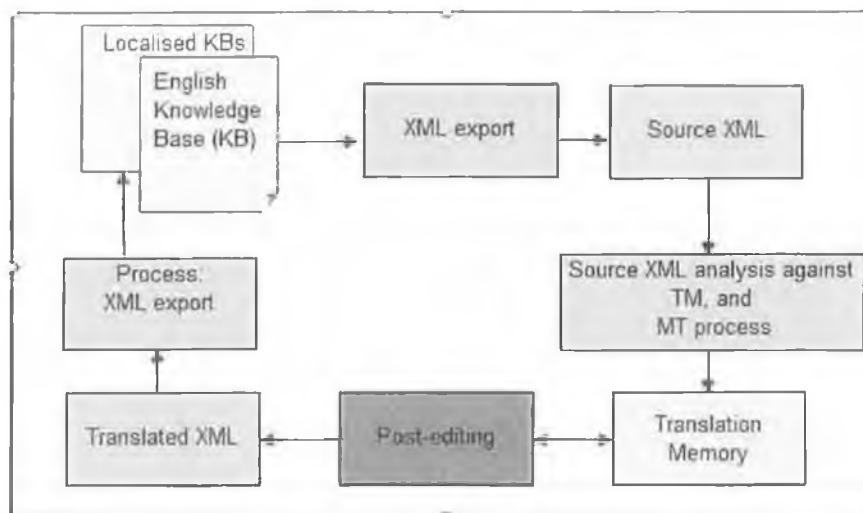
Figure 5-6: Translation workflow using MT

This process could be adapted to set the TM leverage threshold value to 99% so that all source segments that do not have an exact 100% match are sent to the MT engine. The source document would then be pre-translated with the TM, so as to be automatically imported into the knowledge base and published. Such an approach would present the advantage of being close to a real-life 24/7 automatic translation service. Yet, a few issues arise. Firstly, the documents' original content would have to be edited in the knowledge base and automatically exported in an XML format. As discussed earlier, a custom knowledge base template was specifically created for this study. This means that the XML export function would have to be modified to handle this new template, which contains both the controlled and original versions in the same document. Such development work seemed beyond the scope of this study.

Besides, using a large TM would affect the internal validity of the results, since an unknown part of the documents' translation would originate from legacy human translations. It would therefore be impossible to conclude with confidence that the application of CL rules improved the usefulness of MT output since the use of a TM would invalidate the results. One exception to this decision to use MT on its own concerns a CL rule that deals specifically with segmentation issues and reuse. As discussed in the previous chapter, rule 47 states that elements such as titles and error messages should be separated from running text and be used as independent

segments. This type of authoring does not only pose a problem for MT, but it also affects TM reuse, as shown with the following example where the title is in italics:

☒ Follow the steps in *Removing your Norton program using SymNRT* to remove the program

Article titles always appear on their own at the start of a document, so when documents are translated or post-edited, they are stored as independent TM segments. 100% matches can then be retrieved by independent article titles, but not by embedded ones. In this study, the TM component of Systran 5.0 was used exclusively to store the translation of an article name. Apart from this exception, deciding to rely purely on refined MT output meant that Symantec's MT user dictionaries had to be supplemented with any new term occurring in the source documents. The following section deals specifically with the terminological work performed prior to the final MT process.

5.3.2.2. Terminological work performed on the selected documents

Two factors guided the approach taken with the terminology work performed on the English documents: the accuracy requirements outlined in the previous section, and the size of the documents. Since it was decided that all important terms should be included in the terminology extraction process, a manual approach was taken. The approach used was based on the methodology described by Allen (2001: 27), whereby source texts are machine-translated interactively to identify unknown and mistranslated terms.

5.3.2.2.1. Interactive coding of new and unknown terms

The content of each source document was copied and pasted into Systran's interactive translation environment, Systran Translation Project Manager. A project was created for each source document and interactively machine-translated using Symantec's existing user dictionary alongside Systran's Computer/Data processing technical dictionary. After translating both source documents (899 words) a first time, the number of terms originating from Symantec's existing user dictionary was 54 for French and 72 for German. This meant that more new terms had to be encoded for the English-French language pair than for the English-German

language pair. Any term that returned an incorrect translation due to a different context or a missing translation was then manually sent to a trilingual project-specific user dictionary, translated, and encoded, by using Systran coding clues in certain situations. 31 new terms were encoded for French and 19 for German. 10 specific terms were also sent to a reference dictionary to handle User Interface (UI) options by using the lookup operators described in section 4.3.1 of Chapter 4. Figure 5-7 shows some entries containing UI options in the reference user dictionary:

	"Options" (pn)	"\"Options\"" (pn)	"\"Optionen\"" (pn)
	"Scan incoming Email" (pn)	"\"Analyser les emails entrants\"" (pn)	"\"Eingehende E-Mails prüfen\"" (pn)
	"Scan outgoing Email" (pn)	"\"Analyser les emails sortants\"" (pn)	"\"Ausgehende E-Mails prüfen\"" (pn)
	"Permanently" (pn)	"\"En Permanence\"" (pn)	"\"Dauerhaft\"" (pn)

Figure 5-7: Example of reference user dictionary entries

As discussed in section 1.3.4 of Chapter 1, Figure 5-7 shows that common words such as 'Permanently' or 'Options' had to be carefully handled and encoded when they occurred as UI options in the source documents. These entries were marked with double quotation marks for ease of identification in the MT output. Overall, the dictionary coding process took an hour for both language pairs, the most time-consuming task being the looking up of existing translations in reference TMs. Based on previous studies (such as Babych et al., 2004), the impact of this dictionary update work is expected to improve consistently the performance of the MT system. Determining the actual impact of this dictionary update on the reception of MT documents falls outside the scope of this study, but it is essential to highlight that both versions of the documents had their terminology encoded in the user dictionary. Certain exceptions to this process are discussed in section 5.3.2.2.2.

5.3.2.2.2. Exceptions to the coding process

Since some of the CL rules selected for this part of the study operate at the lexical level, not all new words and term variants were coded in Systran's user dictionary. This decision was made to ensure that the effect of these CL rules would be visible in the controlled MT output. The two CL rules concerned were *'Check the spelling'* and *'Omit unnecessary words'*. If spelling mistakes were found in the original documents, they would only be rectified in the controlled documents.

The subjectivity associated with the rule *'Omit unnecessary words'* should not be underestimated, because what seems redundant or unnecessary to one person may sound stylistically legitimate to another. For instance, the phrase 'if for some reason' was used in one of the source documents. After examining the MT output, it transpired that this type of phrase did not return an idiomatic translation, so it could have been possible to add it to the user dictionary as a protected word sequence. Since the meaning of this phrase is already encapsulated in the common conjunction 'if', it was decided not to pollute the user dictionary unnecessarily. The phrase 'for some reason' was then removed from the controlled source document, but left in the original version.

Finally, the guideline concerning the use of consistent vocabulary and terminology was followed in this study. As mentioned in the previous chapter, a project-specific dictionary of generic words was not compiled for this study, but any generic lexical inconsistency in the source documents was addressed in the controlled versions of the documents. For instance, both synonymous phrases "keep somebody from doing something" and "prevent somebody from doing something" appeared in the documents. After translating the source documents interactively, it emerged that the second phrase was handled better by the MT system than the first one, so it was retained for the rewriting of the controlled documents.

This section has presented the methodology that was used to encode new and mistranslated terminology into a specific MT user dictionary, and discussed some of the choices made during the rewriting process. The next section will expand on this rewriting process, by detailing the strategy used to turn original source documents into controlled documents.

5.3.3. Experimental variations

In order to identify rule violations in the two documents, two approaches were considered. First, the subset of CL rules identified in the previous chapter could be formalised and integrated in acrocheck™. The relatively small size of the documents, however, did not justify such development. Besides, some of the rules could not be formalised, so a manual checking and application process was required. Checking rule violations and implementing rewritings was performed when

examining the sample corpus during the interactive MT process using Systran Translation Project Manager. Each instance of rule violation was counted and added to get a total number (#) of violations per document, as shown in Table 5-2.

Document	Number of words (original version)	Number of CL rule violations	Number of words (controlled version)
NAV05_131	325	14	312
NAV05_121	574	18	579

Table 5-2: Number of rule violations and words per document

Table 5-2 shows little increase in word numbers when original documents are turned into controlled documents. This table also highlights the fact that only a limited number of rule violations occurred in the selected documents (the full list of rule violations present in the original versions of these two documents is provided in Appendix H). The CL rules violated in these two documents are presented in Table 5-3:

CL rule name	Violations NAV05_131	Violations NAV05_121
Rule 5: Do not use more than 25 words per sentence.	2	4
Rule 17: Do not use a pronoun when the pronoun does not refer to the noun it immediately follows.	-	1
Rule 24: Repeat the preposition in conjoined prepositional phrases.	-	2
Rule 26: Avoid the ellipsis of verb.	-	1
Rule 28: Omit unnecessary words.	1	3
Rule 36: The subject of a non-finite clause must be the same as that of the main clause.	-	1
Rule 41: Avoid embedded parenthetical expressions introduced by commas or dashes.	1	-
Rule 42: Do not include parenthesized expressions in a segment unless the segment is still valid syntactically when you remove the parentheses while leaving the parenthesized expressions.	-	1
Rule 47: Move document names, error messages, or section titles to independent segments.		3
Rule 49: Do not use "this", "that", "these", or "those" on their own.	-	1
Rule 51: Avoid –ing words.	2	-
Rule 52: Use single-word verbs.	8	1

Table 5-3: Types of CL rules violated in the selected documents

The numbers presented in Table 5-3 suggest that the original documents were not completely uncontrolled, since violations of very effective rules such as rules 19, 22, 35, or 44 were unfortunately not found in the original documents. Only 9 rules from the list of 28 rules presented in Table 4-9 at the end of Chapter 4 were violated in

the two documents selected for the online experiment. Table 5-3 indicates, however, that some of the rules violated in the two documents do not originate from the final list of rules (namely rules 24, 42, and 49). Due to the relative small number of rules violated in the original documents, it was decided to implement three rules with limited impact in order to compensate for the absence of violations of more effective rules. Another strategy may have consisted of the intentional introduction of rule violations in the original versions. But it would have been difficult, if not impossible, to ensure that all CL rules were broken the same number of times. Besides, such an approach would have created an artificial situation that would never be found in a real-life scenario. Some rules failed to be violated because their rewriting overlapped with the reformulation of similar rules. For instance, rule 50, *'Keep both parts of a verb together'* was violated once in the selected documents. The sentence in which the violation occurred was:

☒ This document gives directions to turn email scanning on or off.

Based on this rule, the rewriting should have been:

☑ This document gives directions to turn on or turn off email scanning.

However, the original sentence also violated rule 52, *'Use single-word verbs'*. In this particular example, it seemed preferable to rewrite the sentence based on the second rule, since it was more effective in the first part of the study:

☑ This document gives directions to enable or disable email scanning

Other rules could not be applied systematically. For instance, the above example contains an –ing word that is part of a technical term 'email scanning' so it cannot be reformulated. As mentioned in the previous chapter, it was decided to reduce the scope of rule 51, *'Avoid -ing words'*. In this experiment, the application of this rule was limited to two specific cases: the use of conjoined –ing words and the use of a gerund as subject, as shown in Table 5-4.

Rule violated	Original NAV05_131	Controlled NAV05_131
Rule 51: Avoid –ing words	Turn**ing** on or turn**ing** off email scanning in Norton AntiVirus	**How to** enable or disable email scanning in Norton AntiVirus
Rule 51: Avoid –ing words	**Checking** email for problems is one Norton AntiVirus task.	Norton AntiVirus can scan your email to check for problems.

Table 5-4: Violations of rule 51 in the selected documents

Some of the rules were not violated in the selected documents because they are too specific. For instance, the rule stating that *'A question mark should only be used at the end of a direct question'* proved very effective in the first part of the study due to the segmentation issue it caused. However, the selected documents did not contain a violation of this rule, so it will not be possible to take this rule into account when conclusions are made on the impact of CL rules at the document level. Once two versions were available for each document, they were manually uploaded and formatted in the knowledge base before being published. Whether the small number of modifications introduced in the controlling of original documents is sufficient to show significant differences in users' reactions must be taken into account during the formulation and testing of hypotheses. This will be discussed in the next chapter.

5.4. Summary

This chapter covered the two main steps used to set up an online environment for the second part of the study: the design of an online questionnaire that is embedded within experimental documents, and the preparation of these documents for an MT process. The design of the questionnaire used in this online environment was driven by the requirement to obtain as many responses as possible in a short timeframe. These responses were automatically collected and stored in a database, so as to be exported in a CSV format for analysis. A full analysis and discussion of the results will be presented in the next chapter. The online experiment used two documents containing a controlled version and an original version of the same content, but users were randomly presented with only one of the two versions. The French and German versions of these documents are included in Appendices J and K

respectively. Once analysed, the users' answers will help us determine whether the application of CL rules has any impact on the usefulness, comprehensibility, and acceptability of MT documents even when they are not post-edited.

Chapter 6

Chapter 6: Analysing the impact of CL rules on the reception of MT documents

6.1. Objectives of the present chapter

In the previous chapter, the approach used to implement an online experiment was discussed. The main concern of this chapter is the analysis of the results of the online experiment, by examining the responses submitted by users for two distinct documents. In order to formulate hypotheses about the outcome of these results, the documents used in the online study have been rated by a group of evaluators to obtain a 'quality' benchmark. This group of evaluators is described in detail in the section 6.2.1. The analysis of the data will be performed using both descriptive and inferential statistics. An alpha level of 0.05 or less is adopted for all statistical tests. These statistical results will be discussed to answer the second research question: from a Web user's perspective, does the application of CL rules have any significant impact on the usefulness, the comprehensibility, and the acceptability of documents containing refined MT output? When answering this question, results will be compared between the two locales. The distribution of response rates obtained in this experiment will first be presented. The remainder of the chapter will discuss and compare the answers obtained from users for the two documents.

6.2. Preliminary evaluation of MT documents

6.2.1. Evaluation setup

In order to formulate hypotheses about the effect of CL rules on the characteristics of MT documents, it was decided to use an expert group of independent evaluators. The documents were assessed by asking questions similar to those users were asked to answer. This group of evaluators was composed of four external professional translators in each language. These translators were either freelance translators or employees of translation agencies that Symantec's localisation department uses on a regular basis. These evaluators were asked to answer three questions for each of the document versions they were given to read. In order to avoid any bias with regard to a certain version of the documents, they were not told that the purpose of the experiment was to assess the effectiveness of CL rules on MT output. Also, they were not given the source documents, since the objective of this parallel study was to reproduce the conditions in which Symantec users would operate. Two batches of documents were then randomly created. Each batch comprised one controlled MT document version and one original MT document version so that both versions of a given document were not present in the same batch. The first batch of documents was dispatched with specific instructions, provided in Appendix N. The evaluators were first asked to report on the time it took them to read the document, and then to answer the two following questions:

- Did the quality of the translation hamper your comprehension?

- Would this document be useful for a Symantec user?

Once the first round of evaluation was completed and their answers submitted, they were sent the second batch of documents with the same instructions two days later. They may have noticed that some of the documents were similar to those they had read two days earlier, but they were not asked to compare both versions. Each set of answers was provided individually for each document version[13].

These answers are examined in detail so that hypotheses can be formulated and tested based on the public's responses.

---

[13] The evaluators' answers are included in Appendix M.

6.2.2. Formulating hypotheses

The two documents used in this experiment vary in length and complexity. The original version of NAV05_131 contains 325 words and 2 simple procedural tasks. The document NAV05_121 contains 574 words, describing 2 distinct solutions comprising 3 and 2 tasks respectively. Because of these textual differences, the evaluators' answers were examined separately so as to formulate hypotheses that are specific to a particular document. This approach is used to avoid formulating and testing hypotheses assuming that the experimental materials and conditions are similar. This assumption had been made by Shubert et al. (1995) when testing the effect of SE on the comprehensibility of two documents. After finding differences in their results, they had to re-examine the complexity of the source materials, before concluding that the effects were limited to one procedure (ibid: 173).

6.2.2.1. Evaluation of Document 'NAV05_131'

In the remainder of this chapter, the terms 'Original' and 'Controlled' refer to the two versions of the machine-translated documents used in the experiment. These terms are used to qualify the translated documents, even though CL rules were applied on the source documents. This distinction is similar to the distinction between MT output A and MT output B used in the first part of the study. Based on this distinction, the following answers were obtained from evaluators when they were asked the first question after reading document 'NAV05_131':
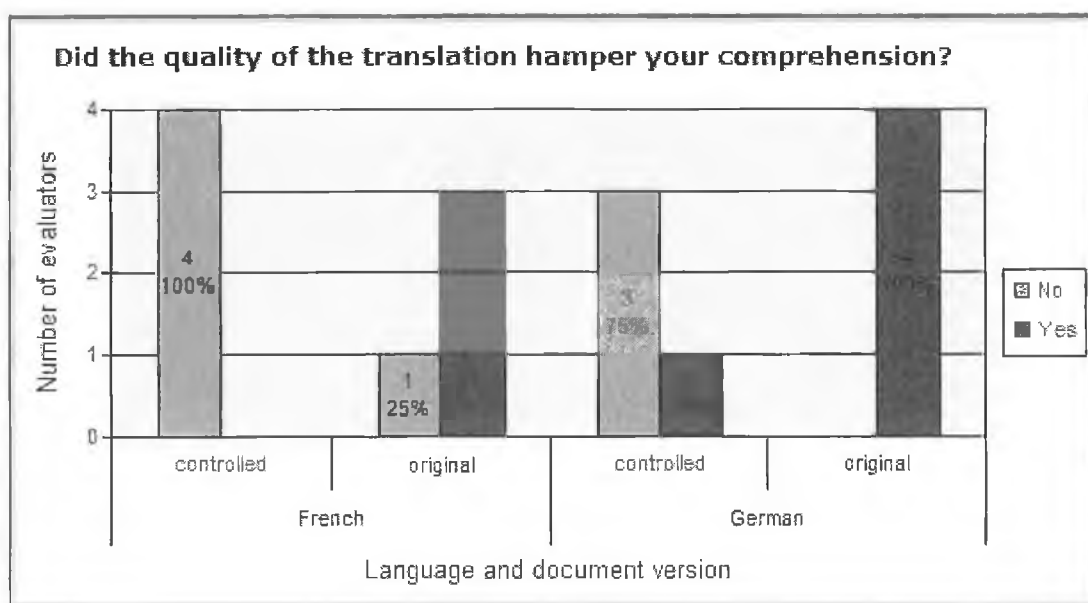

Figure 6-1: Evaluators' first answers for document 'NAV05_131'

The results in Figure 6-1 suggest that the attitude of evaluators on the comprehensibility of the MT output contained in this document is closely related to the use of CL rules for the controlled version of the document. The introduction of CL rules seems to have played its part in removing ambiguity and complexity from source documents, which in turn produced more comprehensible MT output. The four French evaluators thought that the controlled version was easy to understand despite being machine-translated. However, only one of them thought that this was also the case with the original version. The German evaluators were unanimous in finding the original version difficult to understand. When confronted with the controlled version of the document, three out of four German evaluators did not have any major comprehension problem. Based on these results, the following hypothesis can be formulated:

> **H1**: The French and German 'Controlled' versions of document NAV05_131 are more comprehensible than the 'Original' versions. French and German users presented with the 'Original' version of the document are more likely to experience comprehension problems than users presented with the 'Controlled' version.

This hypothesis will be tested during the analysis of the data obtained from users. The following answers were obtained for the second question, 'Would this document be useful for a Symantec user?':
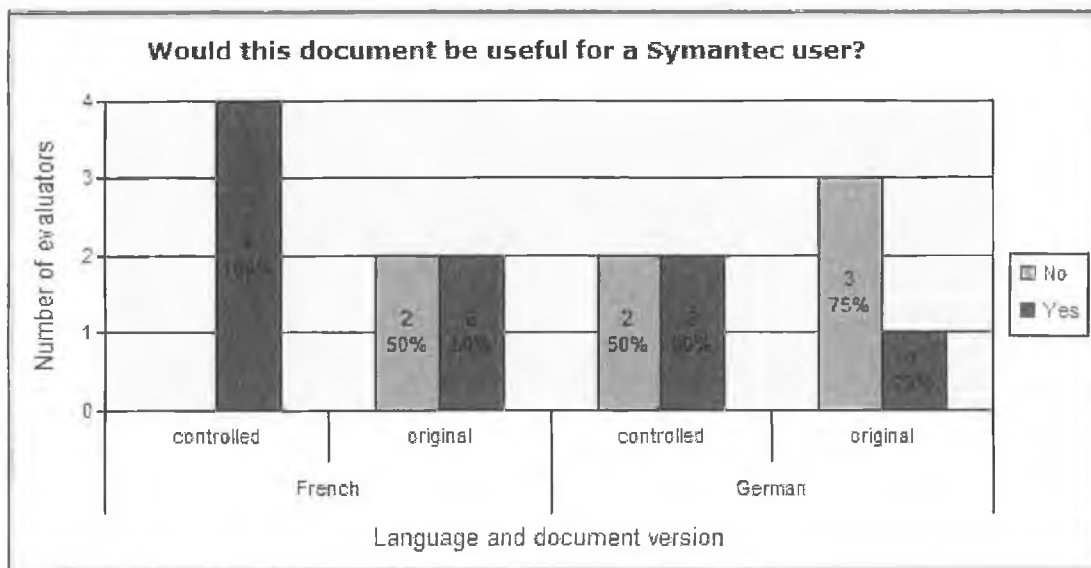

Figure 6-2: Evaluators' second answers for document 'NAV05_131'

Whereas results were relatively consistent between French and German evaluators for the first question, their answers differ for the second question. French evaluators are unanimous in thinking that the 'Controlled' version would be useful for users, whereas German evaluators are divided. Three German evaluators did not report any comprehension problems when answering question 1, but only two of them think that the document would be useful for users. However, the trends remain similar, since a majority of evaluators believe that 'Original' version would not be useful for users.

> **H2**: French and German users presented with the 'Original' version of document NAV05_131 are less likely to find it useful than users presented with the 'Controlled' version.

6.2.2.2. Evaluation of Document 'NAV05_121'

A similar approach was used to formulate hypotheses with regard to certain characteristics of document NAV05_121. The answers to the first question are presented in Figure 6-3:
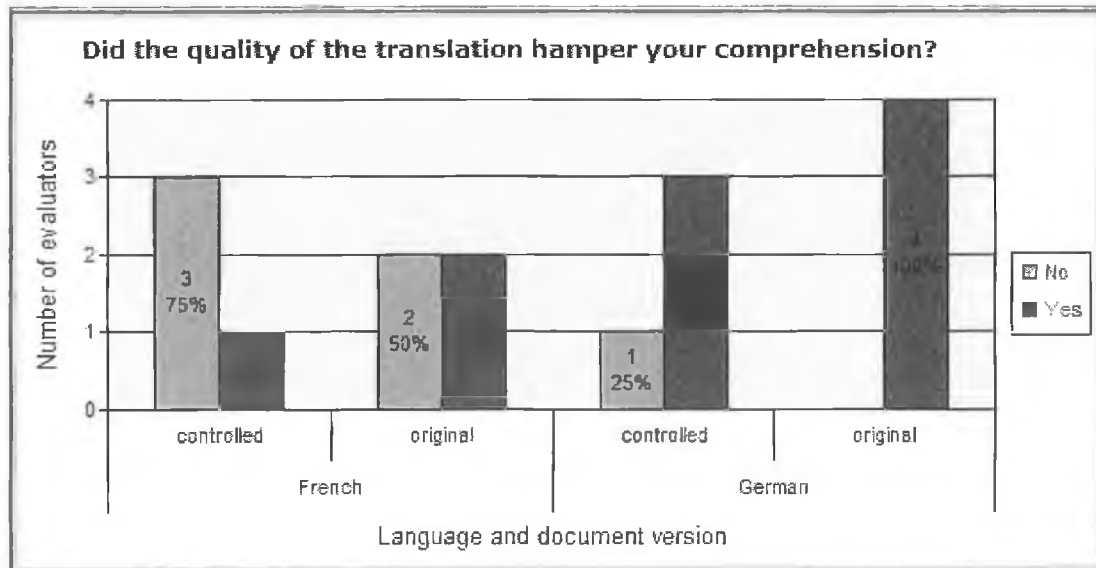

Figure 6-3: Evaluators' first answers for document NAV05_121

The results obtained for the second document, NAV05_131, differ from those obtained for the first document. Results differ between the two locales, since most German evaluators encountered comprehension problems regardless of the document version. In this situation, the effect of CL rules may have been masked by other linguistic phenomena. On the other hand, French evaluators seem to think that the 'Controlled' version is more comprehensible than the 'Original' one. Two different hypotheses therefore have to be formulated:

> **H3A**: The French 'Controlled' version of document NAV05_121 seems to be more comprehensible than the 'Original' version. French users presented with the 'Original' version of the document are more likely to experience comprehension problems than users presented with the 'Controlled' version.

> **H3B**: The German 'Controlled' version of document NAV05_121 does not seem to be more comprehensible than the 'Original' version. German users presented with the 'Original' version of the document are as likely to experience comprehension problems as users presented with the 'Controlled' version.

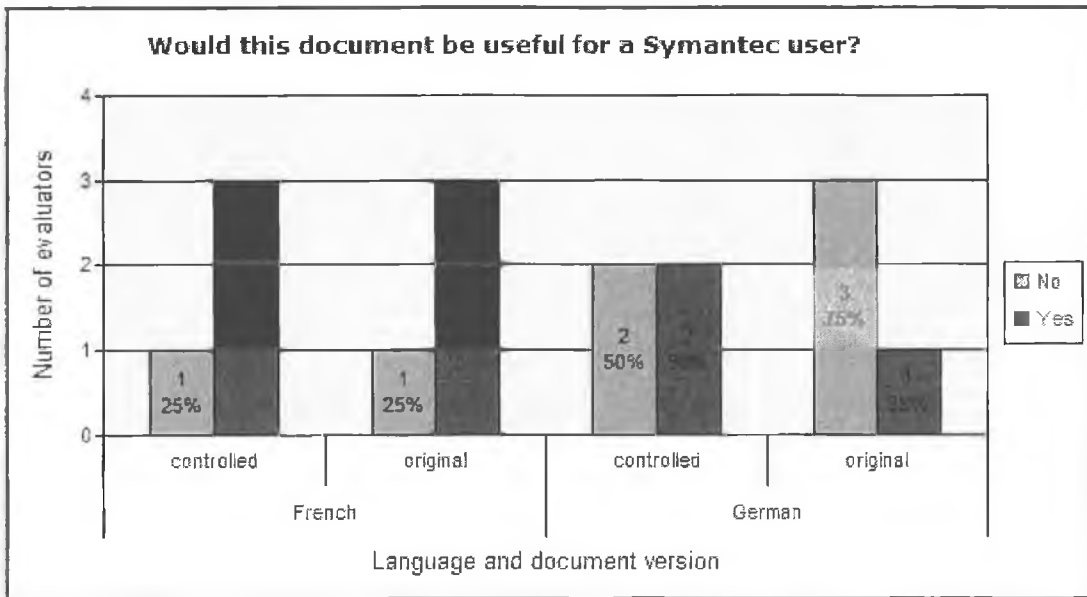The answers obtained for the second question are shown in Figure 6-4:

Figure 6-4: Evaluators' second answers for document NAV05_121

Figure 6-3 showed that two French evaluators encountered comprehension problems with the 'Original' document, but the results presented in Figure 6-4 indicate that three of them believe the document would still be useful for users. This attitude differs from the one observed with German evaluators for the previous document in Figures 6-1 and 6-2. Based on these results, the following hypotheses can be formulated:

H4A: French users presented with the 'Original' version of document NAV05_121 are as likely to find it useful as users presented with the 'Controlled' version.

H4B: German users presented with the 'Original' version of document NAV05_121 are less likely to find it useful than users presented with the 'Controlled' version.

In order to test these hypotheses, the proportions of positive and negative answers obtained from groups of users will be compared based on the document version they consulted. The selection of a statistical test is described in the next section.

6.2.3. Selecting statistical tests to test the hypotheses

Some of the hypotheses (H3 and H4) are formulated differently depending on the locale. The distribution of user responses will therefore be presented per target language (or locale) and document version using four independent groups of respondents. In order to test the formulated hypotheses, response differences between groups will be observed and compared. As mentioned by Levine and Stephan (2005: 137), 'the hypothesis test examines the difference between the proportions of two groups'. Z-tests will therefore be conducted to determine whether the proportion of positive or negative answers obtained for a specific question from a particular group (such as 'French Original') is significantly different from the proportion of positive or negative answers obtained from another group (such as 'French Controlled').

In order to test the hypotheses formulated in the previous section, a significance level must be chosen. De Vaus (2002: 230) states that 'the lower the significance level, the more confident we are that our observed percentage differences reflect real differences in the population'. For dichotomous variables, he suggests using tests using '0.05 for small samples and 0.01 or lower for larger samples' (ibid). Based on the size of the present samples, a level of significance of 0.05 was chosen for the Z-tests.

Sections 6-3, 6-4, 6-5, and 6-6 focus on the analysis of the responses received from users, starting with the public's response rates.

6.3. Analysis of the distribution of responses

6.3.1. Response rates

In the previous chapter, it was mentioned that response rates obtained during the pilot study were relatively low, with 4% of responses from German users and 3% from French users for document NAV05_131. During the month following the pilot study, response rates increased slightly for this particular document, as shown in Table 6-1, in which response rates for document NAV05_121 are also presented:

Document: NAV05_131	German	French
Number of Web hits	1870	1488
Number of responses	85	73
% of respondents	4.70	4.91
Document: NAV05_121	German	French
Number of Web hits	1551	3180
Number of responses	39	54
% of respondents	2.51	1.70

Table 6-1: Response rates (June 6th – July 9th 2006)

Table 6-1 focuses on the response rates obtained after the pilot study. These rates and numbers do not include the 12 German answers and the 8 French answers received for NAV05_131 during the pilot study. These 20 answers, however, will be used for the data analysis.

Table 6-1 suggests that a large number of Symantec users accessed the two documents used in the experiment without answering the questionnaire. Based on the typology of non-respondents established by Bosnjak et al. (2001: 12), these individuals may have been 'lurkers' who read the questions but did not complete the questionnaire. They may also have been individuals who started answering the questionnaire but did not submit it. Determining the proportion of each category of non-respondents is impossible. Besides, Bosnjak et al. (2001: 13) remind us that Web hits may not necessarily be triggered by humans, but possibly by 'robots, worms, or wanderers'. If this were the case, the actual response rates would be higher than the ones presented in Table 6-1.

In section 5.2.4 of Chapter 5, the technical implementation of the online experiment was described by explaining that the randomisation process would rely on storing both versions of the same document in the same knowledge base's location. Due to this implementation, it is not possible to know how many hits were specifically received for the controlled versions or the original versions of a given document. Since both document versions were published within the same knowledge base's location, they did not receive a unique document identifier allowing for the tracking of specific Web hits.

The two documents published for the duration of this study did not have the same exposure. Based on the number of Web hits, document NAV05_121 proved more popular with French users than with German users. This might be due to the technical issue described in the document, which deals specifically with a third-party product (MSN Messenger).[14]

---

[14] The users' results are included in Appendix L.

Response rates also vary greatly between the two documents, but it is impossible to determine with certainty the causes of these differences. Possible explanations may include the length, complexity and the quality of the document. Besides, if users did not consult documents in their entirety or stopped after the description of the first solution in document NAV05_121, they may not have reached the questionnaire located at the bottom of these documents. This assumption seems to be supported by the use of the pop-up mechanism, as shown in Figure 6-5:
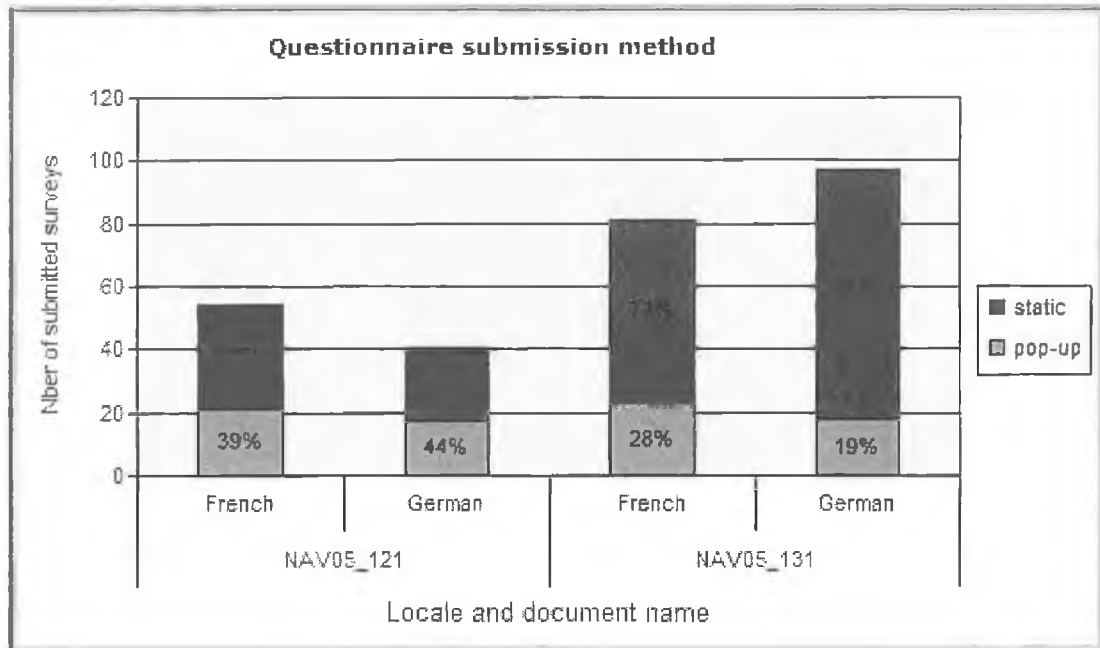

Figure 6-5: Responses submission method

The results show that the pop-up mechanism proved very useful to encourage users to submit feedback, especially with document NAV05_121. To some extent, this usage conflicts with the view expressed by Nielsen and Loranger (2006: 74), who state that 'empirically, we see many users close pop-ups as fast as possible — often even before the content has been rendered. (…) Sad to say, even good pop-ups are rarely appropriate these days, however, because evil pop-ups have tarnished their reputation.' The number of responses submitted via the pop-up in this study does not confirm this statement, since a large proportion of users used this mechanism to submit their feedback for the longer document of the two, NAV05_121. However, it should be mentioned that the pop-up probably did not reach all users since certain Web browsers can be configured to automatically block them.

6.3.2. Distribution of responses based on document version

Since the objective of this study is to compare users' responses based on the document version they were given to consult and their locale, four groups of respondents are used for the analysis of each document. These four groups of respondents are: 'French Controlled', 'French Original', 'German Controlled', and 'German Original'. The distribution of submitted responses based on locale and document version is shown in Figure 6.6:
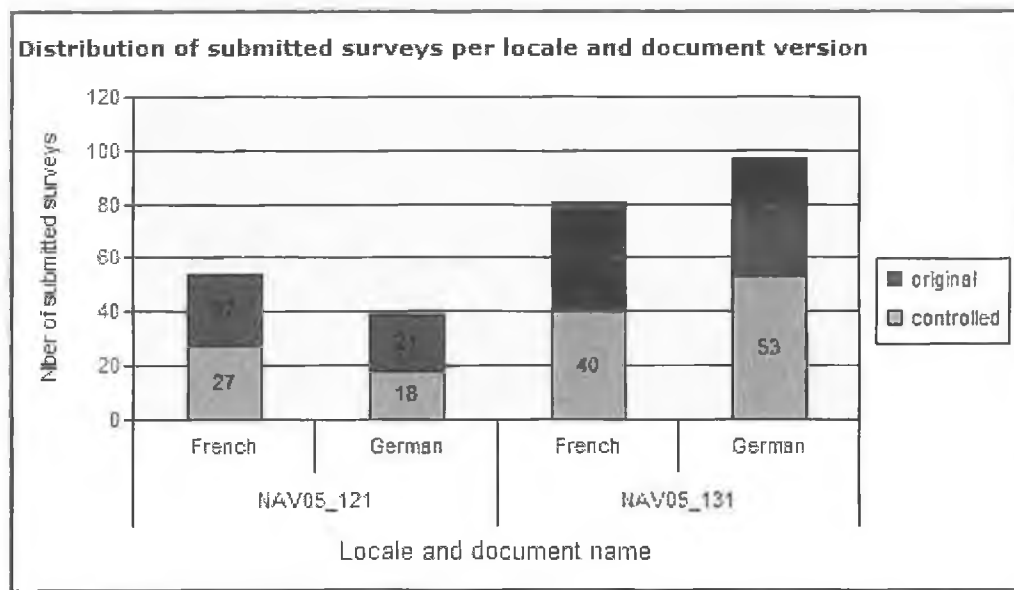

Figure 6-6: Distribution of submitted responses

The above figure shows that the number of responses obtained was almost the same for each type of document version. A total of 133 responses was submitted when the MT document presented to users corresponded to an original source document, which is slightly less than the total of 138 responses received when the MT document originated from the controlled version of a document. The distribution of responses per locale is also homogeneous since 136 responses were obtained from German users, and 135 responses from French users. This distribution of responses is ideal for statistical tests since there is no small cell across the four groups of respondents for each document.

6.4. Analysis of the responses obtained for the first document

The next section focuses on the analysis of the impact of specific CL rules on the comprehensibility and usefulness of the first document, NAV05_131, which contains 14 rule violations in its original version. These rule violations are detailed in Appendix H.

6.4.1. Evaluating the impact of CL rules on the comprehensibility of the first document

In the experiment, users were asked to provide answers after consulting a document. This does not necessarily mean that they read the entire document before giving their answers. For instance, certain users may have decided to focus on a specific part of the document. Since document NAV05_131 contains two procedural tasks, certain users may have read only the first section, the one that describes the task required to enable email scanning. Taking this factor into account, the following responses were obtained from users:
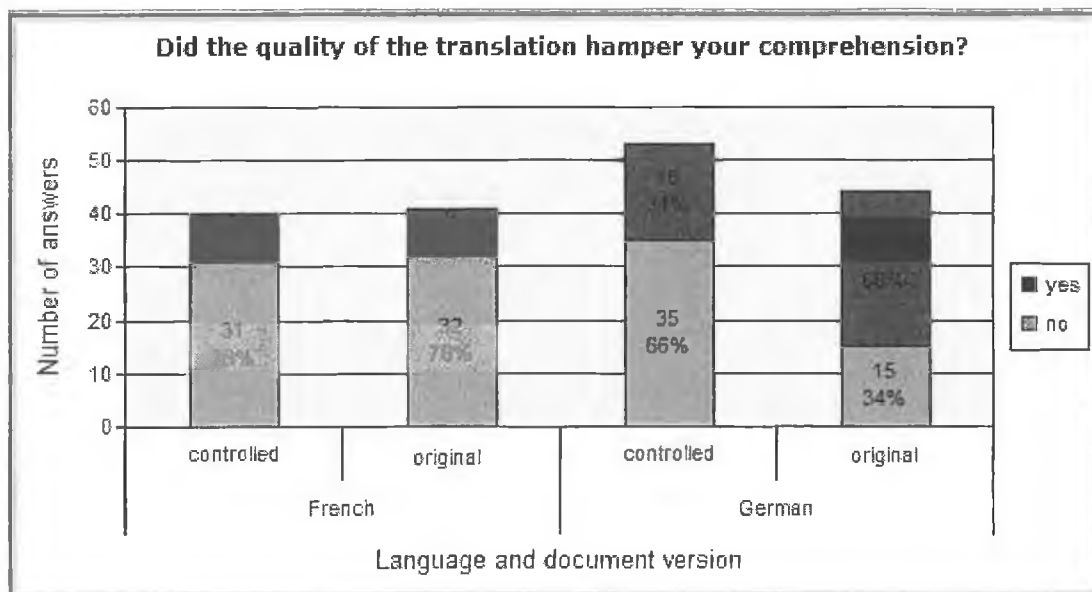

Figure 6-7: Users' answers to the third question for document 'NAV05_131'

Hypothesis H1 is tested by comparing the proportions of negative answers obtained from users to this particular question based on the document version they consulted:

> **H1**: The French and German 'Controlled' versions of document NAV05_131 are more comprehensible than the 'Original' versions. French and German users presented with the 'Original' version of the document are more likely to experience comprehension problems than users presented with the 'Controlled' version.

Results indicate that there is a significant difference in the proportions of negative answers obtained from German users who consulted the Original version of the document and those who consulted the Controlled version. On the other hand, there is no significant difference in the proportions of negative answers obtained from French users who consulted the Original version of the document and those who consulted the Controlled version. Hypothesis H1 is therefore confirmed by the German results, but contradicted by the French results.

In order to provide potential explanations for these results, the MT output contained in the French 'Original' document must be examined to determine whether some of the CL rules were more effective in German than in French. For instance, let us consider the third sentence of the 'Situation' section. In the 'Original' version of the document, this sentence contained a violation of rule 52, stating that verbs with particles should be replaced with single-word verbs where possible.

- ☒ This document gives directions to turn email scanning on or off.
- A. Ce document donne des directions à l'analyse du courrier électronique de tour en fonction ou hors fonction.
- ☑ This document gives instructions to enable or disable email scanning.
- B Ce document donne des instructions pour activer ou désactiver l'analyse du courrier électronique.

The above examples show that in the original sentence, the verb 'turn' and the particles 'on' and 'off' were not analysed properly by the MT engine, which produced French MT output that would traditionally be regarded as incorrect and unintelligible. In MT output A, 'turn' is translated as a noun, and the particles 'on' and 'off' are translated as phrases. In the 'Controlled' version, the ambiguity was removed by the application of the rule, so correct analysis and generation phases were performed by the MT system, resulting in a grammatically correct MT output.

The same problem occurred in the German output:

☒ This document gives directions to turn email scanning on or off.

A Dieses Dokument gibt Richtungen zum Umdrehung E-Mail-Prüfung an oder weg.

☑ This document gives instructions to enable or disable email scanning.

B Dieses Dokument erteilt Anweisungen, E-Mail-Prüfung zu aktivieren oder zu deaktivieren.

Despite finding similar examples of an effective rule application in both French and German output, results differ between the two locales with regard to the overall impact of the rules at the document level.

A potential explanation for differences in the overall impact of CL rules on the comprehensibility of this particular document lies in the textual location of unintelligible sentences. The example mentioned earlier is directly located after the title of the document, in a section whose purpose is to provide background information or an overview of the instructions contained in the 'Solution' section. The 'Situation' section usually contains a few generic sentences, such as:

> Checking email for problems is one Norton AntiVirus task. Email scanning checks each email. This document gives directions to turn email scanning on or off. The directions are for Norton AntiVirus 2003 through 2006.

One may argue that such a section is redundant since most of the information is already contained in the title of a document such as 'Turning on or turning off email scanning in Norton AntiVirus'. For certain users, reading and acting on the instructions contained in the 'Solution' section may be the most critical part of their online technical experience, so they may skim over the 'Situation' section. Whether the sentences of this section are totally intelligible may not be so relevant. Of course, this is only a hypothesis that would need to be confirmed by a more thorough usability study.

This hypothesis, however, is supported by the fact that few rules were broken in the 'Solution' section. Besides, some of the rules violated in this section did not seem to have any impact on the well-formedness of the output. An example of rule 52, which was broken twice in the 'Solution' section, is provided below with the corresponding translated sentences showing no sign of improvement:

- ☒ To turn on email scanning
- ⒜ Pour activer l'analyse du courrier électronique
- ⒜ E-Mail-Prüfung aktivieren
- ☑ To enable email scanning
- ⒝ Pour activer l'analyse du courrier électronique
- ⒝ E-Mail-Prüfung aktivieren

The above example is different from the one provided earlier, because the particle 'on' was correctly analysed by the MT system, probably due to the fact that it immediately followed the verb.

Overall, these results suggest that French users did not regard the document as holistically as their German counterparts, probably focusing on the 'Solution' section. These results may also suggest that French users' tolerance for unintelligible output may be higher as long as their overall online experience results in solving their technical problem. This type of attitude already provides some elements of information with regard to how useful this MT document was to users. This aspect is discussed in section 6.4.2.

6.4.2. Evaluating the impact of CL rules on the usefulness of the first document

The first two questions that users were asked to answer concern the usefulness of the document they were confronted with. Since the answers to the second question could have been influenced by a wide range of variables, the analysis of the results will focus only on the first question. The second question, 'Did the document answer your question?,' was part of Symantec's original customer satisfaction feedback mechanism so it was retained in the questionnaire used in this study. As discussed in section 5.2.3.1, however, this question was not precise enough to determine the usefulness rate of a document. It was expected that more clear-cut

results would be obtained with the question 'Was this document useful?', whose distribution of answers is presented in figure 6-8:
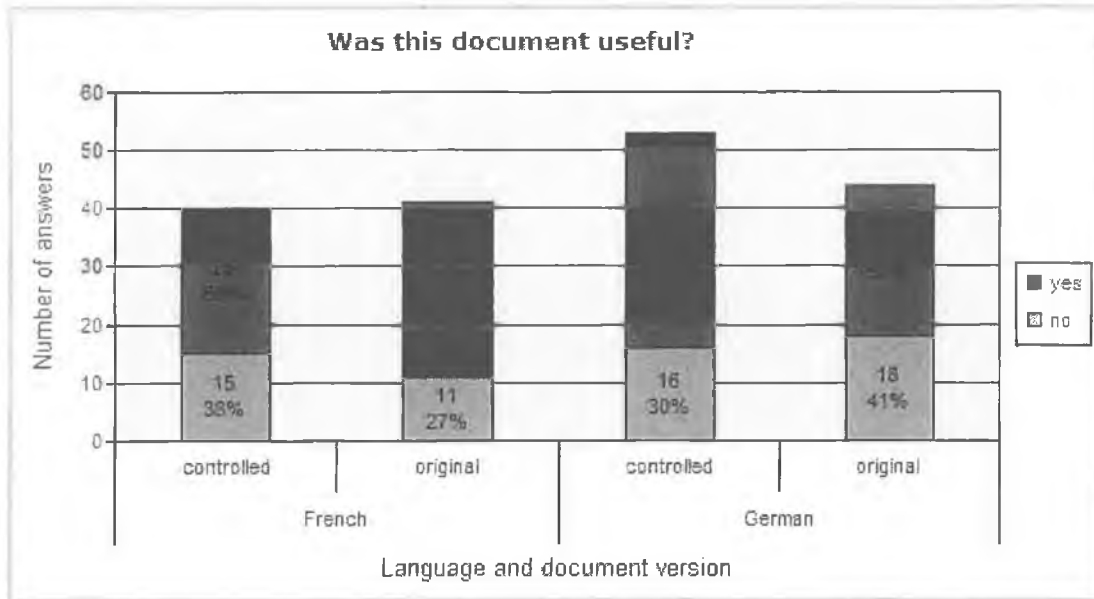

Figure 6-8: Users' answers to the first question for document 'NAV05_131'

The results for question 1 show that in the four groups of respondents, at least 50% of respondents found the document useful. Regardless of the version users were given to consult, these initial results confirm that the refined MT output contained in this document was useful for a majority of users. Hypothesis 2 was then statistically tested:

> **H2**: French and German users presented with the 'Original'
> version of document NAV05_131 are less likely to find it useful
> than users presented with the 'Controlled' version.

Results indicated that there was no significant difference when comparing the proportions of positive answers obtained from the groups 'French Controlled' and 'French Original', and 'German Controlled' and 'German Original'. These results, which contradict the second hypothesis, may be explained by the reason that has been advanced earlier. The most important section of the document is the 'Solution' section, since it contains the steps that must be performed by users to accomplish a task. Very few CL rules were broken in this section, so the machine-translated content present in this section is almost identical in both versions of the document. The usefulness of this section seems to be influenced by the intrinsic simplicity of the sentences used in this section.

The usefulness of this section, and possibly of the whole document, in its 'Original' version may also be influenced by another parameter. This parameter is the correct translation of technical terms in both versions of the documents. As described in Chapter 5, all technical terms, including User Interface (UI) options, were encoded in the MT system's user dictionaries. This pre-processing step was necessary to ensure that these UI options would perform their iconic role, as discussed in section 1.3.2.2.1 of the first chapter. The results obtained for this particular document suggest that this terminological work was sufficient for document NAV05_131 to prove useful for most users, regardless of the application of certain CL rules. This finding must be examined in the light of the results obtained for the second document, NAV05_121.

6.5. Analysis of the responses obtained for the second document

As noted in section 6.2.2, the second document selected for this experiment is more complex than the first one, since it contains two solutions composed of several tasks. The types and number of rules violated in the document NAV05_121 are also different. These rule violations are included in Appendix H. Separate hypotheses were therefore formulated for this document. The impact of the CL rules on the comprehensibility of the machine-translated content is first discussed.

6.5.1. Evaluating the impact of CL rules on the comprehensibility of the second document

As noted in Table 6-1, fewer responses were obtained for this particular document than for document NAV05_131, so cells are smaller. The responses obtained for the third question are presented in Figure 6-9:
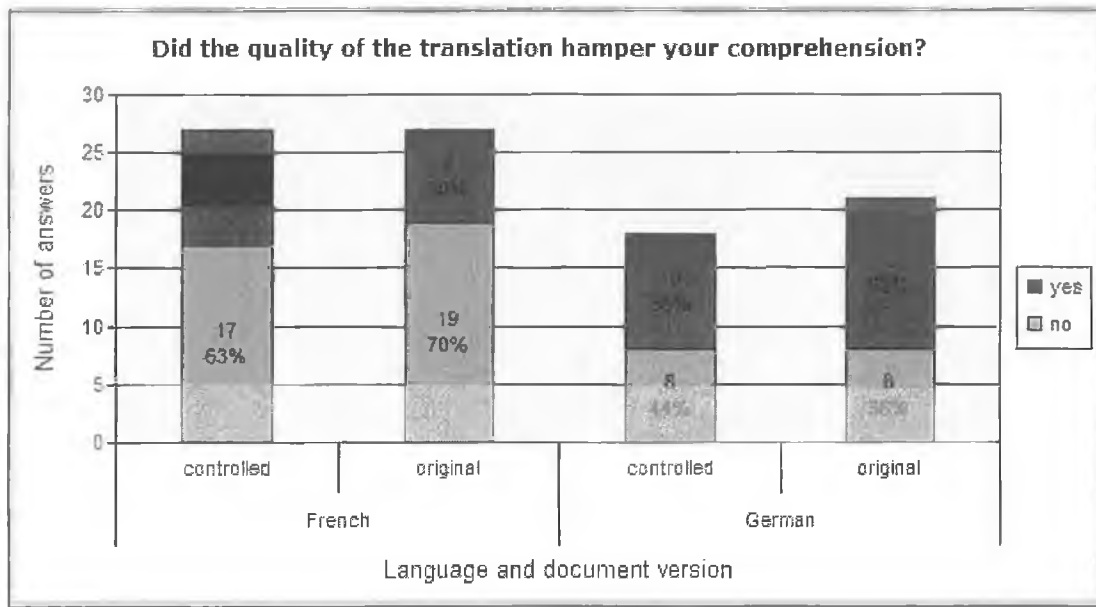

Figure 6-9: Users' answers to the third question for document 'NAV05_121'

The results presented in Figure 6-9 are used to test two separate hypotheses, the first one concerning responses obtained from French users.

> **H3A**: The French 'Controlled' version of document NAV05_121 seems to be more comprehensible than the 'Original' version. French users presented with the 'Original' version of the document are more likely to experience comprehension problems than users presented with the 'Controlled' version.

Results indicate that there is no significant difference between the proportions of negative answers obtained from French users who were presented with an 'Original' version of the document from those who consulted a 'Controlled' version of the document. The high proportion of positive answers obtained from French users who consulted the 'Original' version of the document NAV05_121 may be explained by one of the aforementioned arguments. Certain users may not have read the document in its entirety, and therefore not encountered comprehension problems. For instance, let us examine one of the sentences that violated five CL rules and its translation in the 'Original' version:

☒ If for some reason it is not, continue on to the next section for how to manually configure the latest version of MSN Messenger to work with Norton AntiVirus Instant Messenger scanning.

⚠ Si pour quelque raison elle n'est pas, continuez en fonction à la section suivante pour que la façon configure manuellement la dernière version de MSN Messenger pour fonctionner avec l'analyse de messagerie instantanée de Norton AntiVirus.

In the above example, the translated sentence should have created comprehension problems for users (with the mistranslation of the 'how to' clause), but without detailed feedback, it is difficult to explain why this did not occur. Besides, the question users were asked to answer could only be answered by 'yes' or 'no', so it may have been difficult for them to report on their comprehension difficulties. A scale of values may have provided more detailed results.

To some extent, the same explanations must be advanced to explain the lack of impact of CL rules on the comprehensibility of the German output. This lack of significant difference was noted after testing the second hypothesis for this particular document:

> **H3B**: The German 'Controlled' version of document 'NAV05_121' does not seem to be more comprehensible than the 'Original' version. German users presented with the 'Original' version of the document are as likely to experience comprehension problems as users presented with the 'Controlled' version.

No significant difference was found between the proportions of negative answers obtained from German users who were presented with an 'Original' version of the document and those who consulted a 'Controlled' version of the document. These results show, however, that it was essential to evaluate separately the impact of certain CL rules on the comprehensibility of MT documents. If the results obtained from users had been analysed regardless of the document (NAV05_121 or NAV05_131), errors of Type 1 or 2 could have been made during the analysis. As stated by Kreyszig (1999: 1118), a Type 1 error is to reject a true hypothesis, and a Type 2 error is to accept a false hypothesis. These results also confirm that once certain CL rules are applied together, it is very difficult to determine their individual

impact. Once again, a more detailed qualitative study may provide more in-depth explanations.

6.5.2. Evaluating the impact of CL rules on the usefulness of the second document

The final two hypotheses concern the impact of the CL rules on the usefulness of the document 'NAV05_121'. Two separate hypotheses were tested based on the locale in which the document was consulted. The first hypothesis concerns the impact of the CL rules on the usefulness of the French MT documents. The second hypothesis deals with the impact of the CL rules on the usefulness of the German MT documents:

> **H4A**: French users presented with the 'Original' version of document NAV05_121 are as likely to find it useful as users presented with the 'Controlled' version.

> **H4B**: German users presented with the 'Original' version of document NAV05_121 are less likely to find it useful than users presented with the 'Controlled' version.

These hypotheses were tested by comparing the proportions of positive answers obtained from the groups of users 'French Controlled' and 'French Original', and 'German Controlled' and 'German Original' respectively. The results obtained from these groups of users are shown in Figure 6-10. In both tests, no significant difference was found between the proportions of positive answers obtained from the 'Original' groups and the 'Controlled' groups', so hypothesis H4A must be accepted and H4B rejected.
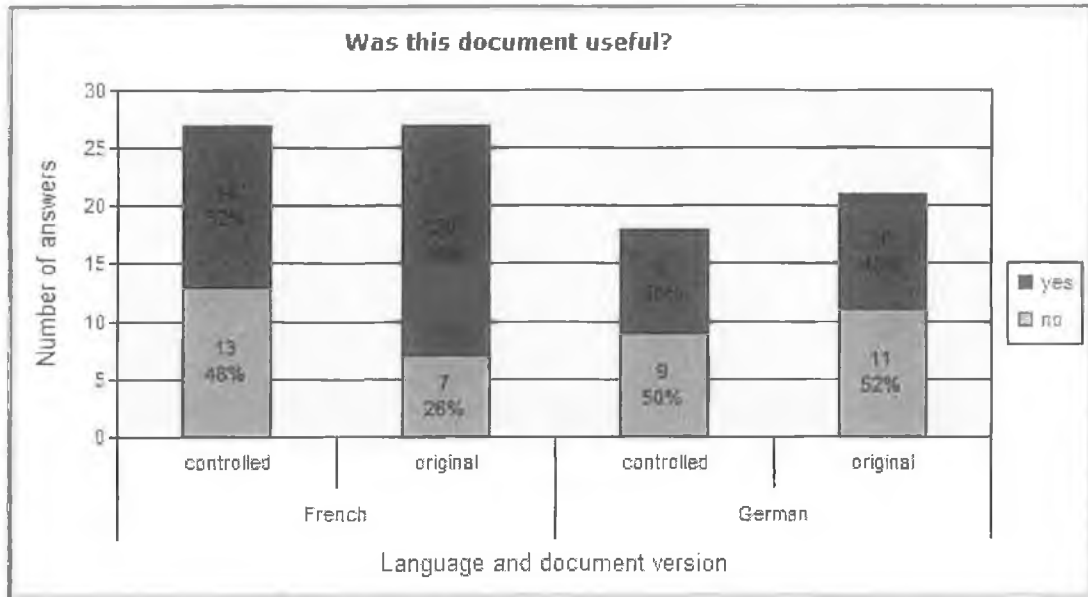
Figure 6-10: Users' answers to the first question for document 'NAV05_121'

These results confirm what was hypothesised after the analysis of the results for document NAV05_121. The introduction of extra CL rules in documents that do not originally contain many CL rule violations is not the most important factor to take into account when analysing the usefulness of such technical support documents. This statement lacks precision, but it seems very difficult to determine accurately when a document becomes uncontrolled, and when it would benefit from the application of extra CL rules so that the usefulness of corresponding MT documents can be significantly improved. Translated documents containing refined MT output appear to be useful for most users, provided that they contain accurate translations of most technical terms, tokens, and UI options.

Based on these results, a final hypothesis must be formulated and tested, so that the final question of this study may be answered: does the introduction of extra CL rules have any impact on the acceptability of MT documents? This hypothesis is tested and discussed in the next section.

6.6. Impact of CL rules on the acceptability of MT documents

As discussed in section 5.2.3.3 of Chapter 5, the concept of acceptability is not easy to grasp so it was decided not to ask this question directly to users. The acceptability of MT documents was evaluated with the following two questions:

- Would you have preferred to read the English version of this document?

- If you were to encounter another technical problem in the future, would you consult again a machine-translated document?

The first of these two questions was asked to check the respondents' need for localised technical information. After all, the original document is available in English, so certain users may prefer to read the English version. However, it has been assumed in this study that users of consumer software products may have a limited knowledge of English. Having access to machine-translated information in their own language might be better than a mere pis aller in certain situations. Figure 6-11 shows the results obtained for the fourth question for both documents:
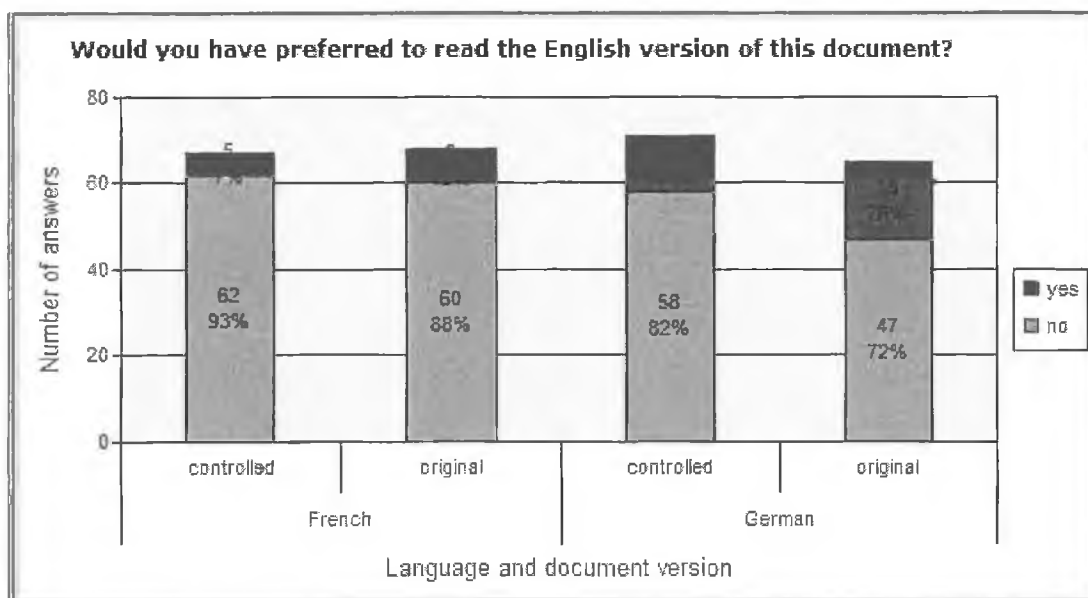

Figure 6-11: Users' answers to the fourth question (both documents)

It was decided to combine both documents for the analysis of these results after checking that there was no significant difference at the document level when comparing the proportions of positive answers provided by two distinct groups. The results presented in Figure 6-11 clearly show that regardless of the document version, an overwhelming majority of users would not have preferred to read the English version of the document. These results suggest that the English skills of most users of Symantec consumer products are insufficient for them to find a solution in original documents. MT documents therefore have the potential to offer access to information they would be unable to get. As noted in the previous section, the final question in this second study must be answered: does the introduction of extra CL rules have any impact on the acceptability of MT documents?

In order to answer this question it was decided to test two hypotheses. The first one follows the model that has been used throughout the analysis of the results:

> **H5**: Users presented with the 'Original' version of a document are less likely to consult again such a document than users presented with a 'Controlled' version

The second hypothesis is prompted by some of the results obtained earlier. If certain users encounter comprehension problems when consulting a machine-translated document, they may be unwilling to consult such documents again in the future. Based on the results obtained during the testing of Hypothesis H1, German users seem to be less tolerant than French users with regard to the quality of MT documents. The following hypothesis is therefore formulated:

> **H6**: German users presented with an 'Original' version of a document are less likely to find this type of document acceptable than French users presented with an 'Original' version

The answers obtained from both documents for the fifth question are presented in Figure 6-12:
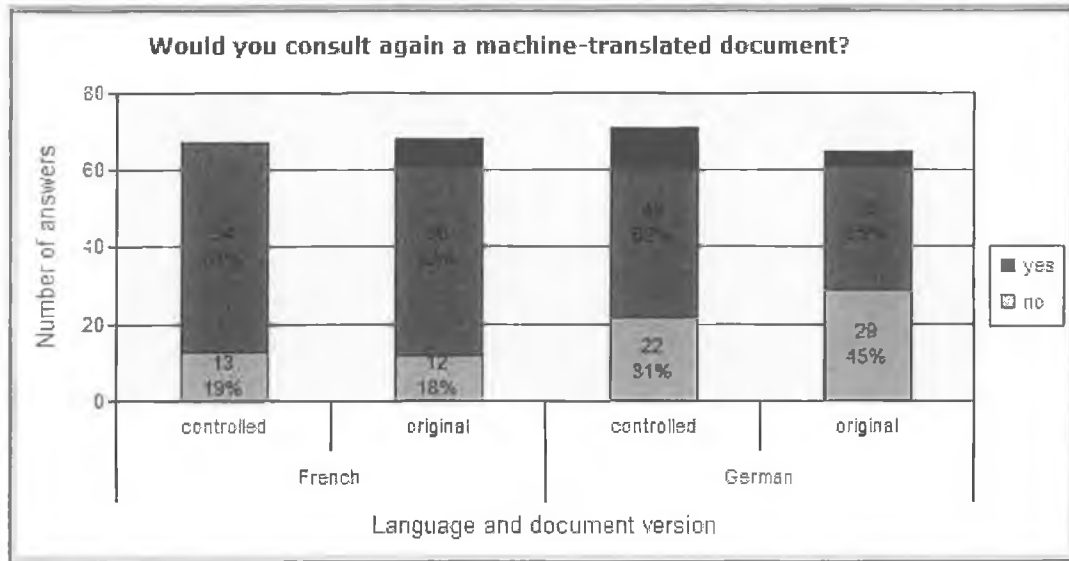

Figure 6-12: Users' answers to the fifth question (both documents)

It was decided to test hypotheses H5 and H6 by combining the answers obtained for both documents after checking that no significant difference existed at the document level. Results, which are presented in Figure 6-12, did not reveal any significant difference when comparing the proportions of positive answers obtained from users presented with an 'Original' version of a document with those obtained from users presented with a 'Controlled' document. Hypothesis H5 therefore cannot be confirmed by these results.

On the other hand, hypothesis H6 was confirmed by the results obtained from French and German users. There is a significant difference between the proportions of positive answers obtained from German users presented with an 'Original' document and those obtained from French users with similar documents. From a customer service perspective, having differences in the way users respond to a type of document is problematic. This problem is amplified by the fact that certain users lack English skills while not finding MT output acceptable enough to consider consulting another MT document in the future. If an MT service was the only service provided, these users or customers might start looking elsewhere for a similar product and service.

Possible explanations for these different attitudes towards MT output have already been provided, one of them being the relative inferiority of the MT system's English-German pair compared to the English-French pair. The difference in word order between the three languages is indeed more likely to penalise German than French. Besides, more tolerance from the disciples of L'Académie Française with regard to an MT version of their own language, even if imperfect, may account for a different distribution of answers across the two locales.

6.7. Data analysis conclusions

One of the main findings of the second part of this study is that some of the hypotheses formulated based on the expert opinions of translators had to be rejected. This was the case for hypotheses H1 for French, H2 for French and German, H3A, and H4B. The lack of agreement between translators and users may be explained by a variety of reasons. First of all, users and translators do not have the same linguistic standards and expectations, so translators may be more likely to be affected by translation inaccuracies than users, especially when these translation inaccuracies are generated by MT systems. Besides, evaluators were asked to read a document in its entirety, whereas users may have consulted only part of a document. Since some of the CL rules were violated in sections that were not required for users to complete a task, their impact may have gone unnoticed.

This finding limits the conclusions that can be derived from this study, since other documents containing different rule violations may have generated different results. The fact that results were not consistent across both documents suggest that many parameters have the potential to mask the impact of certain CL rules. Such parameters include the textual location of the rule violation, the amount of terminological preparation performed prior to the MT process, and the ability of the MT system to leave certain sections untranslated. While the impact of certain rules was proved at the segment level, some of them failed to make an impact at the document level.

The impact of certain CL rules was evaluated at the document level by asking users to provide their opinions on the usefulness, comprehensibility, and acceptability of MT documents. The most important finding of this study is that the introduction of a set of CL rules in source documents produced different results depending on the locale. Whereas no significant difference was noted with the answers of French users for any of the questions, a significant difference was found in the proportion of answers obtained from German users with regard to the comprehensibility of one of the documents. These results clearly suggest that more studies are required using different documents, text types, and language pairs.

Chapter 7

Chapter 7: Conclusion

7.1. The objectives

This research had two main objectives. The first objective was to determine whether certain MT-oriented CL rules would be more effective than others in improving the comprehensibility of French and German MT segments. This objective was met when comparing the results obtained from the evaluation of two sets of MT segments. It must be stressed, however, that improvements sometimes only became visible when performing a micro-analysis of these MT segments. In certain cases, the use of well-defined evaluation and analysis metrics was not sufficient to consistently determine whether a rule was effective. This analysis challenge was explained by a number of reasons which will be summarised again in section 7.2

The second objective of this research was to investigate whether the application of certain CL rules would impact on the reception of specific MT technical documents from a Web user's perspective. This objective was met by conducting an online experiment involving genuine users of Symantec products. This was the first time a study of this type was conducted to evaluate the impact of CL rules on MT documents. Such a study was only made possible thanks to this researcher's privileged position at Symantec. Without the support of Symantec's Technical Support and Localisation groups, such a study could not have involved genuine users in a genuine setting. This unique case study suggests that the use of clearly-defined rules can significantly improve the comprehensibility of German technical support documentation that has been machine-translated but not post-edited. For French, the introduction of CL rules did not improve the usefulness, comprehensibility, and acceptability of MT documents. The findings for German must be interpreted very carefully since the results were not consistent across the two documents that were used in this experiment. Additional findings of this research are summarised in section 7.2.

7.2. Findings

The results of both parts of this study confirm that the comprehensibility of MT output can be improved by the application of CL rules in certain situations to such an extent that no comprehension problems are reported at the segment level and document level. This finding brings empirical evidence to support the claim that refined MT output can sometimes be useful, comprehensible, and acceptable for a majority of users when source content is strictly controlled. In the previous statement, the word 'sometimes' is of paramount importance, because cases have been identified when the impact of certain CL rules is not visible. Our data show that machine-translated technical support documents can sometimes be useful to a majority of users despite containing CL rule violations. This finding is consistent with what was found in studies conducted with a different MT system or different language pairs (Richardson: 2004, Jaeger: 2004).

In section 4.4.9 of Chapter 4, the CL rules that proved the most effective for both French and German at the segment level were identified. Examples of such CL rules include rules that specifically proscribe the use of conjoined verbs when they have different valency, the use of pronouns that have no specific referent, or the use of phrasal verbs. It is our contention that these rules could be used in a systematic way to significantly improve the machine-translatability of Web-based technical support documentation. These rules are a good complement to other very effective rules which are often cited in writing guidelines. These include the use of spell-checked content, grammatical constructions, and standard punctuation.

The analysis of the results obtained in the first part of the study enabled us to make additional discoveries, which are limited to the two language pairs used in this study. In certain cases, the scope of certain rules is sometimes too broad, leading to situations where the application of the rule has no significant effect on the comprehensibility of certain MT segments. For instance, this was the case with the rule prescribing the use of relative pronouns. This was explained by the fact that the MT system can properly handle allegedly complex or ambiguous structures, sometimes keeping the ambiguity in the target language. Individual CL rules can

also have a negative impact on the comprehensibility of MT output, when ambiguity is introduced by certain rewritings.

It was also found that the impact of individual CL rules is sometimes masked by other linguistic phenomena, such as polysemy or homography, which can only be handled using a controlled lexicon. Finally, certain individual rules are effective in one language pair only, because the MT system's analysis and transfer rules differ from one language pair to the next.

At the document level, results differ between the two language pairs, since the introduction of extra CL rules significantly improved the comprehensibility of one German MT document whereas it did not lead to any significant difference with the French documents. These results suggest that another factor, the use of fully updated user dictionaries, plays a significant role in ensuring the usefulness, comprehensibility, and acceptability of MT documents. Further research is required to determine the precise impact of this parameter with regard to the role played by the application of CL rules.

Our data also show that regardless of the use of CL rules, an overwhelming majority of users is likely to favour a machine-translated document instead of its English version. These results therefore present a good opportunity for MT services to be provided for this type of content as long as key parameters are respected (such as the use of customised user dictionaries and simple sentences). If these key parameters were not respected, dissatisfied users who cannot benefit from MT content might look for information elsewhere. In this situation, the *raison d'être* of a translated document would be undermined, as mentioned by Loffler-Laurian (1996: 86):

> La traduction est un objet fourni à un client qui doit en être
> satisfait pour en consommer d'avantage.

This trend was noticeable with German users. The present study found that German users are less likely than French users to consult MT documents again when source content is not controlled. This difference in attitude may be explained by genuine translation issues contained in German MT documents, or by the relatively higher linguistic standards expected by German users. Further research is required to determine whether these differences can be reproduced with other language pairs based on the methodological framework used in this study.

7.3. Review of the methodology used

Due to the lack of empirical studies in the field of CL and MT, it was decided to combine two distinct experiments to evaluate the impact of CL rules at the segment and document levels. If rules had only been evaluated at the segment level, it would have been difficult to draw any conclusions on their overall impact. A CL rule may prove very effective in improving the comprehensibility of individual MT segments, but its impact may be limited at the document level if the rule is rarely violated in original documents. Likewise, if rules had only been evaluated as part of a rule set at the document level, it would have been impossible to determine their individual contribution. This two-phase approach, despite certain weaknesses reviewed below, is an original contribution to the methodology of CL evaluation mainly due to the number of MT-oriented rules being evaluated. The test suite used in the first part of this study could therefore be used as a starting point for future studies. As stated in sections 4.4.5 and 4.4.6 of Chapter 4, the wide scope of certain rules coupled with the number of diverse reformulations associated with specific rules prompts for more investigation into the impact of sub-rules. This CL evaluation framework could be easily re-used due to its modularity. In this study, rules have been evaluated using a large number of examples, focusing on the output produced by one MT system into two languages. The output produced by other MT systems in other target languages could be evaluated in a similar fashion. Similarly, the impact of these rules on stylistic characteristics of the source text could also be investigated.

During the first evaluation phase, the size of the test suite used was both a strength and a weakness. To our knowledge, a test suite of this size had never been employed in an empirical study of CL rules using human evaluators. In order to make this study as thorough as possible and make a breakthrough in the field of CL

evaluation, a large number of rule violations was collected. In order to evaluate the effectiveness of each of these MT-oriented rules, diverse linguistic patterns were collected to avoid making hasty conclusions. From an evaluation perspective, mixing examples from sets A and B at random would have presented several advantages. First, it would have prevented evaluators from sensing that certain segments had undergone some modifications and were part of a specific set of items, be it controlled or uncontrolled. This would have ensured that they did not feel obliged to give a better score to one of the sets. Second, it may have removed potential reliability issues with regard to the scores attributed to the set containing MT output B. Since evaluators were asked to score both sets separately, starting with MT output A, their attitude towards MT output may have changed after completing the evaluation of the first set. As mentioned by Krings (2001: 11), their internal criteria may have changed during the evaluation of this first set, so they may have started the evaluation of MT output B with different eyes, being accustomed to some of the output produced by the MT system. Mixing the items of both sets would have prevented this situation. Using a random mix of items is the strategy that was used in the second part of the study when asking a group of evaluators to rate two versions of two machine-translated documents. This strategy, which was described in section 6.2.1 of Chapter 6, seems appropriate for future studies in the field of CL evaluation, be it at the segment or document level.

Besides, due to the final size of the test suite, the evaluation process proved cumbersome for evaluators, as shown by certain scoring inconsistencies. During the data analysis, these scoring inconsistencies were removed by using a combination of median scores and integrity checks. Asking unmotivated users to conduct the same evaluation process at the segment level would not have been possible. This was confirmed by the lack of response obtained in the second part of the study when users massively failed to respond to five short questions.

One of the weaknesses of the online experiment performed in the second phase of this study was indeed the low response rates it generated. Since response rates were below 4%, one can only speculate as to what the rest of the users thought. As mentioned by Lassen (2003: 81), it is not possible to 'provide evidence that those who responded are representative of those who did not respond'. To our knowledge,

no previous study had used such a methodological framework to evaluate certain characteristics of translated materials. It was therefore difficult to predict that users would be so reticent in providing feedback. Despite this weakness, asking genuine users to give their opinions on certain characteristics of MT documents allowed us to determine that users' answers did not always match those provided by evaluators. This finding makes a great contribution to the methodology of CL evaluation, showing that genuine users must be involved in empirical studies focusing on the role of CL rules. Relying purely on translators/evaluators does not appear sufficient to make claims on the impact of certain rules on MT output, since the evaluators' bias towards MT as a technology is a factor that must not be neglected. Different results may have been obtained if 'fake' users (such as translation students) had been employed instead. Having access to genuine users, even if their answers were scarce, was therefore a great advantage to explore new horizons. It also enabled us to establish that the difficulties involved in the evaluation of translations by users should not be underestimated.

Another potential weakness of this online experiment concerns the number of response categories included in the questionnaire. The selection of dichotomous responses was motivated by the need to avoid 'too many fine distinctions' (De Vaus, 2002: 107) that could have confused the respondents and possibly impacted the response rates. One of the issues associated with the approach taken in the present study, however, is that these dichotomous responses did not capture the respondents' real position with regard to the usefulness or comprehensibility of the documents – for instance, the documents may have been completely useless and incomprehensible to some users, but to others, somewhat useful and comprehensible. In future studies, a larger set of ordered responses may be considered once the consequences of having non-committal answers have been taken into account. Using a three-valued scale may provide more fine-grained answers, as long as it does not trigger 'sitting on the fence answers' (De Vaus, 2002: 106). Future studies may yield interesting results if respondents were asked to qualify how comprehensible MT documents are (easy to understand, difficult to understand, or impossible to understand).

On the other hand, one of the strengths of this study lies in the experimental variations that were performed to generate controlled documents out of original documents. The aim of this study was to work with genuine original materials so as to avoid creating artificial documents that may never be produced in a real life scenario. The original documents chosen in this study contained violations of a limited number of rules. To some extent, this lack of rule violations was initially perceived as a disappointment, because a thorough evaluation of 54 CL rules had been performed in the first part of the study. The absence of rule violations in the original documents meant that the impact of certain CL rules would not be evaluated at the document level unless rule violations were artificially introduced. If such an approach had been taken to turn original documents into 'uncontrolled' documents, the translation differences between 'uncontrolled' and controlled MT outputs may have been more significant. However, the external validity of the study would have been undermined because such 'uncontrolled' documents may never be produced by professional content developers.

The results obtained in this study confirmed this assumption, showing that the impact of extra CL rules may not be significant when key CL rules are (unconsciously) being adhered to by content developers. This statement may be refined by hypothesising that the rules that are adhered to in a specific part of a given document are more important than the rules that are violated in less essential sections of a document. Future studies in the field of machine translatability would need to take this challenge into account. If one wanted to determine a confidence level for the machine-translatability of a document and the subsequent usefulness or comprehensibility of its MT output, both rule adherence and rule violations should be taken into account. Other challenges encountered during this study and requiring further research are discussed in section 7.4.

7.4. Future challenges and research opportunities

The findings of the present study are limited to one MT system with two language pairs, and one text type with two documents. The results of the second part of the study showed that the introduction of a set of CL rules did not have any significant impact on the usefulness, comprehensibility, and acceptability of two French MT documents while significantly improving the comprehensibility of one German MT document. Based on these findings, language pairs could be diversified in future studies to determine whether results obtained with Romance languages such as Spanish and Italian are similar to those of French. On the other hand, an Asian language with a different word order to English, such as Japanese, and traditionally high linguistic standards might produce results closer to those obtained for German. Future research could also involve different text types and MT systems.

The findings of the second part of this study are limited to the perceived usefulness of MT documents by users. In order to determine whether CL rules would have any impact on the actual usability of these documents, a genuine usability study would have to be conducted in a controlled environment. Such a study may provide some explanations for the discrepancies found between the evaluators' answers and the users' answers during the data analysis. Based on the data obtained in this study, certain users do not seem to be affected by 'incorrectly' translated sentences. Future studies could further investigate the translation problems that have a negative impact on users or certain categories of users. In this study, users were not asked to provide any examples of sentences that hampered their comprehension so as to avoid burdening them with detailed feedback. In a controlled environment, asking for and getting this type of feedback may be done more easily.

The results obtained in both parts of this study also show that a set of CL rules is sometimes not sufficient to significantly improve the comprehensibility of MT documents. If a fully automated process were to be used in a production environment, the use of CL rules may need to be combined with other automated processes. In section 3.3.1.2.1, the concept of automated post-editing was discussed. This concept has been described by Allen and Hogan (2000) and Allen (2003), but

few studies have explored automated post-editing in a systematic way (Knight & Chander 1994, Elming, 2006). Yet, certain customised MT systems, such as the SPANAM system used at the Pan American Health Organization, make use of macros to automate some of the post-editing tasks (Aymerich, 2001: 5). More research is therefore required to investigate whether certain PE replacements could be completely automated using a bilingual environment. Such automatic replacements could be useful to avoid implementing CL rules that are rarely effective. To some extent, a similar approach may also be used to automatically rewrite source text to make it more machine-translatable without asking authors to adhere to a rule that is not always effective. Such an approach has been described in Shirai et al. (1998) and Nyberg et al. (2003: 276). However, more studies are required to determine the improvements generated by these automatic replacements, when they are performed prior to the MT process as a pre-processing step of after the MT process as a post-processing step.

7.5. Final words

This study has demonstrated that despite using strictly-defined evaluation criteria determining accurately the effectiveness of a particular CL rule is a complex process. While it is possible to formulate some conclusions on the impact of certain rules at the segment level, generalising these results at the rule set and document levels is complicated by several parameters. Before embarking on the strict implementation of a set of CL rules in a particular environment, the impact of these rules should therefore be fully evaluated. Besides, one should not forget that the design of specific CL rules may preclude the refinement of existing MT systems. Certain linguistic phenomena may currently not be handled accurately by specific MT systems in certain language pairs, so a drastic approach would be to develop CL rules to restrict the usage of these phenomena. Such an approach is restrictive and does not seem to encourage future investment in the research and development of MT technology. If the CL and MT paradigm were to be used in more diverse situations, a balance must be found between the implementation of key CL rules and the refinement of certain components of MT systems.

References

References

Acrolinx, 2005. *Acrocheck™ version 2.6.* Berlin: Acrolinx.

Adams, A.H., G. Austin, and M. Taylor. 1999. 'Developing a Resource for Multinational Writing at Xerox Corporation'. In *Technical Communication*, Volume 46, Number 2. pp. 249-254.

Adriaens, G. 1996. 'SECC: Using Test Structure Information to Improve Checker Quality and Coverage'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 226-232.

Adriaens, G. and L. Macken. 1995. 'Technological evaluation of a controlled language application: precision, recall, and convergence tests for SECC'. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI95), Vol.1.* Centre for Computational Linguistics, Leuven, Belgium. pp. 123-141.

Adriaens, G. and D. Schreurs. 1992. 'From COGRAM to ALCOGRAM: Toward a Controlled English Grammar Checker'. In *COLING 92 Proceedings of the 14th International Conference on Computational Linguistics, Vol. III,* Nantes, France. pp. 595-601.

Akis, J. W., and W.R. Sisson. 2002. 'Improving Translatability: A Case Study at Sun Microsystems'. In *Globalization Insider* XI/4.5. 4 pp. Available at: http://www.lisa.org/archive_domain/newsletters/2002/4.5/akis.html?printerFriendly =yes [Last accessed: 18/12/2003].

Allen, J. 2003. 'Post-Editing'. In H. Somers (ed.) *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company. pp. 297-317.

Allen, J. 2003. 'Controlled Translation: the Integration of Controlled Language (CL) and Machine Translation (MT)'. Panel talk presented at the European Association for Machine Translation and Controlled Language Applications Workshop (EAMT/CLAW2003). 15-17 May 2003. Available at: http://www.ctts.dcu.ie/allen.ppt [Last accessed: 04/04/2005].

Allen, J. 2001. 'Postediting: an integrated part of a translation software program'. In *Language International*, April 2001, Vol. 13, No. 2. pp. 26-29.

Allen, J. and C. Hogan. 2000. 'Toward the Development of a Postediting Module for Raw Machine Translation Output: a Controlled Language Perspective'. In *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000),* Seattle, Washington. pp.62-71.

Almqvist, I., and A. Sågvall Hein. 1996. 'Defining ScaniaSwedish – a Controlled Language for Truck Maintenance'. In *Proceedings of the First Controlled Language*

Application Workshop (CLAW 1996), Centre for Computational Linguistics, Leuven, Belgium. pp. 159-165.

ALPAC. 1966. *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, Washington, D.C.

Amant, K.S. 2003. 'Designing effective writing-for-translation intranet sites'. In *IEEE Transactions on Professional Communication,* Volume 46, Issue 1, March 2003. pp. 55-62

Armistead, C.G, and G. Clark. 1992. *Customer Service and Support: Implementing Effective Strategies.* London: Pitman Publishing.

Arnold, D. 'Why Translation is Difficult for Translators'. 2003. In H. Somers (ed.) *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company. pp. 119-142.

Arnold, D., L. Balkan, R. Lee Humphreys, S. Meijer, and L. Sadler. 1994. *Machine Translation: An Introductory Guide.* Manchester: NCC Blackwell.

Aymerich, J. 2001. 'Generation of Noun-Noun Compounds in the Spanish-English Machine Translation System SPANAM'. In *Proceedings of MT Summit VIII,* Santiago de Compostela, Spain. 6 pp.

Babych, B., D. Elliott, and A. Hartley. 2004. 'Extending MT evaluation tools with translation complexity metrics'. In *COLING 2004 Proceedings of 20th International Conference on Computational Linguistics,* University of Geneva, Switzerland, August 2004. pp. 106-112.

Baker, K. L., A. M. Franz, P. W. Jordan, T. Mitamura, and Eric H. Nyberg. 1994. 'Coping With Ambiguity in a Large-Scale Machine Translation System'. In *COLING 94 Proceedings of the 15th International Conference on Computational Linguistics,* Kyoto, Japan. pp. 90-94.

Balkan, L. 1994. 'Test Suites: Some Issues in Their Use and Design'. In *Machine Translation: Ten Years On, Proceedings of the second international conference on Machine Translation organised by Cranfield University and the Natural Language Translation Specialist Group of the British Computer Society (BCS-NLTSG).* British Computer Society: 9 pp.

Barthe, K. 1996. 'Eurocastle – A User's Experience With Prototype AECMA SE Checkers'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 42-63.

Barthe, K., C. Juaneda, D. Leseigneur, J-C. Loquet, C. Morin, J. Escande, and A. Vayrette. 1999. 'GIFAS Rationalized French: A Controlled Language for Aerospace Documentation in French'. In *Technical Communication,* Volume 46, Number 2, May 1999. pp. 220-229

Bernth, A. 2006. 'EasyEnglishAnalyzer: Taking Controlled Language from Sentence to Discourse Level'. Paper presented at the Fourth Controlled Language Application Workshop (CLAW 2006), Boston Marriott, Cambridge, MA, USA. 7 pp.

Bernth, A. 1999. 'Controlling Input and Output of MT for Greater Acceptance'. In *Proceedings of the 21st ASLIB Conference, Translating and the Computer.* London, 1999. 10 pp.

Bernth, A. 1998. 'EasyEnglish: Preprocessing for MT'. In *Proceedings of the Second International Workshop on Controlled Language Applications (CLAW 1998),* Pittsburgh, Pennsylvania. pp. 30-41.

Bernth, A. and M.C. McCord. 2000. 'The effect of Source Analysis on Translation Confidence' In J. White (ed.) *Envisioning Machine Translation in the Information Future: Proceedings of the 4th Conference of the Association for MT in the Americas, AMTA 2004,* Cuernavaca, Mexico. LNAI 1934, Springer-Verlag: Berlin Heideiberg. pp. 89-99.

Bernth, A. and C. Gdaniec 2001. 'MTranslatability'. In *Machine Translation,* December 2001, Vol. 16, No. 3. Dordrecht: Kluwer Academic Publishers. pp. 175-218.

Bloor, T. and M. Bloor. 2004. *The Functional Analysis of English: A Hallydayan Approach* (Second Edition). London: Arnold.

Bosnjak, M., T. Tuten, and W. Bandilla. 2001. 'Participation in Web Surveys – A Typology. *ZUMA Nachrichten* 48. pp. 7-17.

Bowker, D., and Dillman, D. 2000. 'An Experimental Evaluation of Left and Right Oriented Screens for Web Questionnaires'. Paper presented at the Annual meeting of the American Association for Public Opinion Research, May 2000. 19 pp.

Bowker, L. and J. Pearson 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Brants, T. 2000. 'TnT – A Statistical Part-Of-Speech Tagger'. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000,* Seattle, WA.

Bredenkamp, A., B. Crysmann, and M. Petrea. 'Building Multilingual Controlled Language Performance Checkers'. In *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000),* Seattle, Washington. pp. 83-89.

Bryman, A. and D. Cramer. 2001. *Quantitative Data Analysis with SPSS release 10 for Windows.* Hove: Routledge.

Burke, B. 2004. *Worldwide Antivirus 2004-2008 Forecast and 2003 Vendor Shares*. IDC report 31737. Available for purchase at: http://www.idc.com/getdoc.jsp?containerId=31737. [Last accessed: 03/09/005].

Byrne, J. 2004. *Textual Cognetics and the Role of Iconic Linkage in Software User Guides*. PhD thesis. Dublin City University, Dublin, Ireland.

Cardey, S., P. Greenfield, and X. Wu. 2004. 'Designing a Controlled Language for the Machine Translation of Medical Protocols: The Case of English to Chinese'. In R.E. Frederking and K.B. Taylor (eds.) *Proceedings of the 6th Conference of the Association for MT in the Americas, AMTA 2004, Washington DC, USA*. LNAI 3265, Springer-Verlag: Berlin Heideiberg. pp. 37-47.

Clémencin, G. 1996. 'Integration of a CL-Checker in a Operational SGML Authoring Environment'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 32-41.

Coughlin, D. 2003. 'Correlating Automated and Human Assessments of Machine Translation Quality'. In *Proceedings of MT Summit IX*, New Orleans, USA. pp. 63-70.

D'Agenais, J., and J. Carruthers. 1985. *Creating effective manuals.* Cincinnati: South-Western Pub. Co.

Dabbadie, M., A. Hartley, M. King, K. J. Miller, W. M. El Hadi, A. Popescu-Belis, F. Reeder, and M. Vanni. 2002. 'A Hands-on Study on the Reliability and Coherence of Evaluation Metrics'. In *Proceedings of the Workshop on Machine Translation Evaluation, LREC 2002.* Las Palmas de Gran Canaria, Spain, May 2002. pp. 8-16.

Danks, J.H., and J. Griffin. 1997. 'Reading and Translation'. In J.H. Danks, G.M. Shreve, S.B. Fountain, and M.K. McBeath (eds). *Cognitive Processes in Translation and Interpreting.* London: Sage Publications.

Dirven, R. and M. Verspoor. 1998. *Cognitive Exploration of Language and Linguistics.* Amsterdam: John Benjamins Publishing Company.

Doddington, G. 2002. 'Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics'. In *Proceedings of the Second International Conference on Human Language Technology.* San Diego, CA. pp. 138-145.

Dray, S.M. 2004. 'Usable in New York, Usable in Nairobi'. In *Multilingual Computing & Technology.* Vol. 15, Issue 5. pp. 31-32.

De Beaugrande, R., and W. Dressler. 1981. *Introduction to Text Linguistics.* New York: Longman.

De Preux, N. 2005 'How Much Does Controlled Language Improve Machine Translation Results'. In *Proceedings of the 27th ASLIB Conference, Translating and the Computer.* London, 2005. 14 pp.

De Vaus, D. 2004. *Surveys in social research.* London: Routledge.

Dillman, D., R. Tortora, J. Conradt, and D. Bowker. 1998. 'Influence of Plain vs. Fancy Design on Response Rates for Web Surveys'. *Paper presented at annual meeting of the American Statistical Association,* Dallas, TX. 6 pp.

Douglas, S., and M. Hurst. 1996. 'Controlled Language Support for Perkins Approved Clear English (PACE)'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 93-105.

Dyson, M. and Hannah, J. 1987. 'Towards a Methodology for the Evaluation of Machine-Assisted Translation Systems'. In *Computers and Translation,* 2 (3). pp. 163-176.

Eclipse Foundation. 2005. *Eclipse Platform 3.1.* US: Eclipse Foundation.

Elliott, D., Hartley, A. and E. Atwell. 2004. 'A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation'. In R.E. Frederking and K.B. Taylor (eds.) *Proceedings of the 6th Conference of the Association for MT in the Americas, AMTA 2004, Washington DC, USA.* LNAI 3265, Springer-Verlag: Berlin Heideiberg. pp. 64-73.

Elming, J. 2006. 'Transformation-based correction of rule-based MT'. In *Proceedings of the 11th Annual Conference of the European Association for Machine Translation, Oslo University, Norway.* pp. 219-226.

Esselink, B. 2003. 'Localisation and Translation'. In H. Somers (ed.) *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company. pp. 67-86.

Esselink, B. 2001. 'Web Design: Going Native'. In *Language International.* Vol. 13, No. 1. pp. 16-18.

European Association of Aerospace Industries, 2001. *AECMA simplified English.* CD-ROM. UK: AECMA.

Evans, M., C. Wei, and J. Spyridakis. 2004. 'Using Statistical Power Analysis to Tune-Up a Research Experiment: a Case Study'. In *Proceedings of IPCC 2004, Professional Communication Conference, 2004.* pp 14-18.

Farrington, G. 1996. 'AECMA Simplified English: an Overview of the International Aircraft Maintenance Language'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp 1-21.

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, Maryland, USA. pp. 65-72.

Fouvry, F. and L. Balkan. 1996. 'Test Suites for Controlled Language Checkers'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 179-192.

Freeman, B. 2006. 'What Content Belongs in a CMS?' In *Multilingual Computing & Technology's Content Management Getting Started Guide,* April/May 2006, p. 4.

Fuchs, N.E. and R. Schwitter. 1996. 'Attempto Controlled English (ACE)'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 124-136.

Fukuoka, W., Y. Kojima, and J. Spyridakis. 1999. 'Illustrations in User Manuals: Preference and Effectiveness with Japanese and American Readers'. In *Technical Communication*, Volume 46, Number 2, May 1999, pp. 167-176.

Galesic, M. 2006. 'Dropouts on the Web: Effects of Interest and Burden Experienced During an Online Survey'. In *Journal of Official Statistics,* Vol. 22 No. 2, pp. 313-328.

Gdaniec, C. 1994. 'The Logos Translatability Index'. In: *Technology Partnerships for Crossing The Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in The Americas.* Columbia, Maryland. pp. 97–105.

Gerson, S.J., and S.M. Gerson. 2000. *Technical Writing: Process and Product.* Upper Saddle River: Prentice Hall.

Gillham, W.E.C. 2000. *Developing a Questionnaire.* London: Continuum International Publishing Group.

Godden, K.1998. 'Controlling The Business Environment for Controlled Language'. In *Proceedings of the Second Controlled Language Application Workshop (CLAW 1998),* Pittsburgh, Pennsylvania. pp. 185-189.

Godden, K., and L. Means. 1996. 'The Controlled Automotive Service Language (CASL) Project'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 106-114.

Göritz, A. S. 2006. 'Incentives in Web Studies: Methodological Issues and a Review'. In *International Journal of Internet Science*, Vol. 1, No. 1, pp. 58-70.

Govyaerts, P. 1996. 'Controlled English, Curse or Blessing? A User's Perspective'. In In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 137-142.

Gough, N. 2005. *Example-based machine translation using the marker hypothesis.* PhD Thesis. Dublin: Dublin City University.

Hajič, J., P. Hornola, and V. Kubon. 2003. 'A Simple Multilingual Machine Translation System'. In *Proceedings of MT Summit IX*, New Orleans, USA. pp. 157-164.

Halliday, M.A.K and Hasan, R. 1976. *Cohesion in English.* London: Longman.

Hammerich, I., and C. Harrison. 2002. *Developing Online Content: The Principles of Writing and Editing for the Web.* Toronto: John Wiley & Sons, Inc.

Hartley, A., and C. Paris. 2001. 'Translation, controlled languages, generation'. In E. Steiner, C. Yallop (Eds) *Translation and Multilingual Text Production.* pp. 307-325. Berlin : Mouton de Gruyter.

Hayes, P., S. Maxwell, and L. Schmandt. 1996. 'Controlled English Advantages for Translated and Original English Documents'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 84-92.

Heald, I., and R. Zajac. 1996. 'Syntactic and Semantic Problems in the Use of a Controlled Language'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 205-215.

Hoft, N.L. 1995. *International Technical Communication: How to export Information about High Technology.* New York: John Wiley & Sons, Inc.

Holmback, H., S. Shubert, and J. Spyridakis. 1996. 'Issues in Conducting Empirical Evaluations of Controlled Languages'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 166-177.

Hovy, E., M. King, and A. Popescu-Belis. 2002. 'An Introduction to MT evaluation'. In *Workbook of the LREC 2002 Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics.* Las Palmas de Gran Canaria, Spain. pp. 1.7.

Huijsen, W.O. 1998. 'Controlled Language: An Introduction'. In *Proceedings of the Second Controlled Language Application Workshop (CLAW 1998),* Pittsburgh, Pennsylvania. pp. 1-15.

Hutchins, J. 2004. 'Machine translation and computer-based translation tools: what's available and how it's used'. In Bravo, J. M. (ed.) *A new spectrum of translation studies,* Issue 4. Valladolid: Universidad de Valladolid. 20 pp.

Hutchins, J. 2003. 'Machine Translation: General Overview'. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press. pp. 501-511.

Hutchins, W. J. and H. Somers. 1992. *An Introduction to Machine Translation*. London, Academic Press.

IAI, 2002. *CLAT factsheet*. Available at: http://www.iai.uni-sb.de/docs/clatfactsheet.pdf. [Last accessed: 03/09/2005].

Isabelle, P. 1987. 'Machine Translation at the TAUM Group'. In M. King (ed.) *Machine Translation Today: The State of the Art*. Edinburgh: Edinburgh University Press. pp. 247-277.

ITR Ltd. 2002. *BlackJack 2.03*. Available for purchase at http://www.itrblackjack.com. [Last accessed: 24/01/2006].

Jackob, N, and T. Zerback. 2006. 'Improving Quality by Lowering Non-Response: A Guideline for Online Surveys'. Paper presented at the WAPOR Seminar 'Quality Criteria in Survey Research VI', Cadenabbia, Italy, 2006.

Jaeger, P. 2004. 'Putting Machine Translation To Work - Case Study: Language Translation At Cisco'. Presentation given at the LISA Global Strategies Summit, Crowne Plaza Hotel, Foster City, California, USA, June 21-24, 2004.

Jufrasky, D., and J. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, New Jersey: Prentice Hall.

Kamprath, C., E. Adolphson, T. Mitamura, and E. Nyberg. 1998. 'Controlled Language for Multilingual Document Production: Experience with Caterpillar Technical English'. In *Proceedings of the Second Controlled Language Application Workshop (CLAW 1998)*, Pittsburgh, Pennsylvania. pp. 51-61.

Kaplan, R. N. 2003. 'Syntax'. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press. pp. 70-90.

Kennedy, G. 1998. *An introduction to corpus linguistics*. London: Longman.

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*. pp. 189-203.

King, M., and K. Falkedal. 1990. 'Using Test Suites in Evaluation of Machine Translation'. In *COLING-90 Proceedings of the 13th International Conference on Computational Linguistics*, Helsinki, Finland. pp. 211-216.

Kirkman, J. 1992. *Good Style: Writing for Science and Technology*. London: E & FN Spon.

Kittredge, R. 2004. 'Sublanguages and Controlled Languages'. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics*. New York: Oxford University Press. pp. 430-447.

Kittredge, R. 1982. 'Variation and Homogeneity of Sublanguages'. In Kittredge, R. and J. Lehrberger (eds.) *Sublanguage: Studies of Language in Restricted Semantic Domains*. New York: Walter de Gruyter. pp. 107-137.

Krings, H.P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Koby, G.S (ed.). Kent, Ohio: The Kent State University Press.

Klare, G. 1963. *The Measurement of Readability*. Iowa State University: Ames.

Knight, K. and I. Chander. 1994. 'Automated Postediting of Documents'. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI)*, Seattle, Washington. pp. 779-784.

Kohl, J.R. 1999. 'Improving Translatability and Readability with Syntactic Cues'. In *TechnicalCOMMUNICATION*, Vol. 46, No. 2. pp. 149-166.

Kreyszig, E. 1999. *Advanced Engineering Mathematics*. Singapore: John Wiley and Sons, Inc.

Lalaude M., V. Lux, and S. Regnier-Prost. 1998. 'Modular Controlled Language Design'. In In *Proceedings of the Second Controlled Language Application Workshop (CLAW 1998)*, Pittsburgh, Pennsylvania. pp. 103-113.

Lassen, I. 2003. *Accessibility and Acceptability in Technical Manuals*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Lauscher, S. 2000. 'Translation Quality Assessment: Where can Theory and Practice Meet? In *The Translator: Evaluation and Translation, Special Issue,* Vol. 6, No. 2. pp. 149-168.

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 *COLING 96 Proceedings of the 14th International Conference on Computational Linguistics*, Copenhagen, Denmark, August 1996. pp. 711-716.

Lehrberger, J. and L. Bourbeau. 1988. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. Amsterdam: John Benjamins Publishing Company.

Levin, R. 2005. 'Tools for Multilingual Communication'. In *Multilingual Computing & Technology,* March 2005, Vol. 16, Issue 2. pp. 45-49.

Levine, D.M., and D. F. Stephan. 2005. *Even You Can Learn Statistics.* Upper Saddle River, NJ: Pearson Prentice Hall.

Lieske, C., C. Thielen, M. Wells, and A. Bredenkamp. 2002. 'Controlled Authoring at SAP'. In *Proceedings of Translating and the Computer 24, ASLIB,* London, 2002. (no page numbers).

Loffler-Laurian, A.M. 1996. *La Traduction Automatique.* Lille: Presses Universitaires du Septentrion.

Lux, V., and E. Dauphin. 1996. 'Corpus Studies: a Contribution to the Definition of a Controlled Language'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 193-204.

Marlow, A. 1995. *Technical Documentation.* Oxford: NCC Blackwell.

Marlow, A. 2005. *Quality Control for Technical Documentation.* Titchmarsh: Marlow Durndell.

McCarthy, A. 2005. *An Investigation into the Effect of Controlled Language Rules on the Machine Translation of Multiple Language Pairs.* MA thesis. Dublin City University, Dublin, Ireland.

Microsoft Corporation. 1998. *The Microsoft manual of style for technical publications.* 2nd Edition. Redmond: Microsoft Press.

Mitamura, T. Nyberg, E, and J. Carbonell. 1991. 'An Efficient Interlingua Translation System for Multi-lingual Document Production'. In *Proceedings of the Third Machine Translation Summit.*

Mitamura, T. 1999. 'Controlled Language for Multilingual Machine Translation'. In *Proceedings of Machine Translation Summit VII,* Singapore, 13-17 September 1999. pp. 46-52.

Mitchell P., B. Santorini, and M. Marcinkiewicz; 1993. 'Building a Large Annotated Corpus of English: The Penn Treebank'. In *Computational Linguistics,* June 1993, Vol. 19, Number 2. pp. 313-330.

Møller, M.H. 2003. 'Grammatical Metaphor, Controlled Language and Machine Translation'. In *Proceedings of EAMT-CLAW-03,* Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 95-104.

Moore, C. 2000. 'Controlled Language at Diebold, Incorporated'. In *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000),* Seattle, Washington. pp.51-61

Myerson, C. 2001. 'Beyond the hype'. In *Language International,* Vol. 13, No. 1. pp. 12-14.

Nielsen, J. 2000. *Designing Web Usability: the Practice of Simplicity'.* Indianapolis: New Riders Publishing.

Nielsen, J., and H. Loranger. 2006. *Prioritizing Web Usability.* Berkeley: New Riders Press.

Nirenburg, S. 1987. 'Knowledge and Choices in Machine Translation'. In S. Nirenburg (ed.) *Machine Translation: Theoretical and Methodological issues.* Cambridge: CUP. pp. 1-21.

Nyberg, E., and T. Mitamura. 1996. 'Controlled Language and Knowledge-Based Machine Translation: Principles and Practice.' In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 74-83.

Nyberg, E., T. Mitamura, and W.O. Huijsen. 2003. 'Controlled Language for Authoring and Translation'. In H. Somers (ed.) *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company. pp. 245-281.

O'Brien, S. 2006. *Machine Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis.* PhD thesis, Dublin City University, Ireland.

O'Brien, S. 2003. 'Controlling Controlled English: An Analysis of Several Controlled Language Rule Sets'. In *Proceedings of EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 105-114.

Ó Broin, U. 2005. 'Taking Care of Global Business: Translatability and localization of e-business applications'. In Multilingual Computing & Technology. Vol. 16, Issue 5. pp. 35-39.

Ogden, C.K. 1930. *Basic English: A General Introduction with Rules and Grammar.* London: Paul Treber.

O'Hagan, M. and D. Ashworth 2002. *Translation-Mediated Communication in a Digital World: Facing the Challenges of Globalization and Localization.* Clevedon: Multilingual Matters.

Olohan, M. 2004. *Introducing Corpora in Translation Studies.* New York: Routledge.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. 'BLEU: A Method for Automatic Evaluation of Machine Translation'. In *Proceedings of the 40th Annual Meeting of ACL,* Philadelphia, PA. pp. 311-318.

Petrits, A. 2001. 'EC Systran: The Commission's Machine Translation System'. Brussels: Translation Service, Directorate of Resources and Language Support. Available at: http://europa.eu.int/comm/translation/reading/articles/tools_and_workflow_en.htm [Last accessed: 01/03/2006]. 32 pp.

Pool, J. 2006. 'Can Controlled Languages Scale to the Web?' Paper presented at the Fourth Controlled Language Application Workshop (CLAW 2006), Boston Marriott, Cambridge, MA, USA. 12 pp.

Popescu-Belis, A., M. King, and H. Benantar. 2002. 'Towards a Corpus of Corrected Human Translations.' In *Workbook of the LREC 2002 Workshop on Machine Translation Evaluation: Human Evaluators Meet Automated Metrics.* Las Palmas de Gran Canaria, Spain. pp. 17-21.

Popescu-Belis, A., P. Estrella, M. King, and N. Underwood. 2005. 'Towards Automatic Generation of Evaluation Plans for Context-based MT evaluation'. ISSCO Working Paper 64, August 2005. Available at: http://www.issco.unige.ch/publications/working-papers/papers/femti-issco-wp64.pdf [Last accessed: 22/06/2006].

Power, R., D. Scott, and A. Hartley. 2003. 'Multilingual Generation of Controlled Languages'. In *Proceedings of EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 115-123.

Price, J., and H. Korman. 1993. *How to Communicate Technical Information: A Handbook of Software and Hardware Documentation.* Redwood City: The Benjamin/Cummings Publishing Company, Inc.

Pullman, B. 2006. 'Size Really Doesn't Matter'. In *LISA Globalization Insider,* August 2006. Available at: http://www.lisa.org/globalizationinsider/2006/08/size_really_doe.html [Last accessed: 23/09/2006]

Pym, A. 2004. *The Moving Text: Localization, Translation, and Distribution.* Amsterdam: John Benjamins Publishing Company.

Pym, P. J. 1990. 'Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system'. In Pamela Mayorcas (ed.), *Translating and the Computer 10: The Translation Environment 10 Years on*, London: ASLIB. pp. 80-96.

Quirk, R., Greenbaum, S., Leech, G., and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language '.* London: Longman.
Raman, M, and S. Sharma. 2004. *Technical Communication: Principles and Practice.* Oxford: Oxford University Press.

Ramirez Polo, L., and J. Haller. 2005. 'Controlled Language and the Implementation of Machine Translation for Technical Documentation.' In *Proceedings of the 27th ASLIB Conference, Translating and the Computer.* London, 2005. 8 pp.

Reeder, F. 2004. 'Investigation of Intelligibility Judgments'. In R.E. Frederking and K.B. Taylor (eds.) *Proceedings of the 6th Conference of the Association for MT in the Americas, AMTA 2004, Washington DC, USA.* LNAI 3265, Springer-Verlag: Berlin Heideiberg. pp. 227-235.

Reid-Smith, E. 2000. *E-Loyalty: How to keep customers coming back to your website*. New York: Harper Collins Publishers.

Reips, U.D. 2002. 'Standards for Internet-based experimenting'. In *Experimental Psychology,* Volume 49, Issue 4. pp. 243-256.

Reuther, U. 2003. 'Two in One: Can it work? Readability and Translatability by Means of Controlled Language'. In *Proceedings of EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 124-132.

Rice, C. 1997. *Understanding Customers*. Oxford: Butterworth-Heinemann.

Richardson, S., Dolan, W., Menezes, A., Pinkham, J. 2001. Achieving commercial quality translation with example-based methods. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain. pp. 293-298.

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.) *Proceedings of the 6th Conference of the Association for MT in the Americas, AMTA 2004, Washington DC, USA*. LNAI 3265, Springer-Verlag: Berlin Heideiberg. pp. 246-251.

Rochford, H. 2005. *An Investigation into the Impact of Controlled Language across Different Machine Translation Systems: Translating Controlled English into German*. MA thesis. Dublin City University, Dublin, Ireland.

Ross, C. 2004. *Chinese Grammar*. New York: McGraw-Hill.

Roturier, J. 2004. 'Assessing Controlled Language rules: Can they improve performance of commercial Machine Translation systems?' In *Proceedings of the 26th ASLIB Conference, Translating and the Computer*. London, 2004. (no page numbers).

Roturier, J., S. Kramer, and H. Düchting. 2005. 'Machine Translation: The Translator's Choice'. *Presentation given at the 10th Annual Internationalisation and Localisation Conference, LRC X, The Global Initiative for Local Computing*, Limerick, Ireland.

Rychtyckyj, N. 2006. 'Standard Language at Ford Motor Company: A Case Study in Controlled Language Development and Deployment'. Paper presented at the Fourth Controlled Language Application Workshop (CLAW 2006), Boston Marriott, Cambridge, MA, USA. 10 pp.

Rychtyckyj, N. 2002. 'An Assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company'. In S. Richardson (ed.) *Machine translation: from Research to Real Users, Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002*, Tiburon, CA, USA, October 8-12. LNAI 2499, Springer-Verlag: Berlin Heideiberg. pp. 207-215.

Sager, J.C. 1994. *Language Engineering and Translation: Consequences of Automation.* Amsterdam: John Benjamins Publishing Company.

Schachtl, S. 1996. 'Requirements for Controlled German in Industrial Applications', In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 143-149.

Schäfer, F. 2003. 'MT Post-Editing: How to Shed Light on the 'Unknown Task'. Experiences made at SAP'. In *Proceedings of EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 133-140.

Schmidt-Wigge, A. 1998. 'Grammar and Style Checking in German'. In *Proceedings of the Second Controlled Language Application Workshop (CLAW 1998),* Pittsburgh, Pennsylvania. pp. 76-86.

Schwertel, U. 2000. 'Controlling Plural Ambiguities in Attempto Controlled English (ACE)'. In *Proceedings of the Third International Workshop on Controlled Language Applications (CLAW 2000),* Seattle, Washington. pp.105-119.

Scott, M. 2004. *WordSmith Tools 4.0.* Oxford: Oxford University Press.

Senellart, J., P. Dienes, and T. Váradi. 2001. 'New Generation Systran Translation System'. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain. 6 pp.

Senez, D. 1998. 'The Machine Translation Help Desk and the Post-Editing Service'. In *Terminologie & Traduction 1-1998, OPOCE*, Brussels: European Commission. pp. 289-295.

Shackel, B. 1990. 'Human Factors and Usability'. In Preece, J. & L. Keller (eds.) *Human-Computer Interaction.* Hemel Hempstead: Prentice Hall International. pp. 27-41.

Shirai, S., S. Ikehara, A. Yokoo, and Y. Ooyama. 1998. 'Automatic Rewriting Method for Internal Expressions in Japanese to English MT and Its Effects'. In *Proceedings of the Second Controlled Language Application Workshop (CLAW 1998),* Pittsburgh, Pennsylvania. pp. 62-75.

Shubert, S.K, Spyridakis, J.H, Holmback, H.K, and Coney, M.B. 1995. 'The Comprehensibility of Simplified English in Procedures'. In *Proceedings of the Professional Communication Conference:. 'Smooth sailing to the Future'*, IEEE International 27-29 Sept. 1995. p. 171-173.

Slatin, J.M. and S. Rush, 2003. *Maximum Accessibility: Making your Web Site More Usable for Everyone'.* Boston: Addison-Wesley.

Smart. 2005. *Smart Communications.* Available at: http://www.smartny.com/download.htm. [Last accessed: 03/09/2005].

Smart, J. 2006. 'SMART Controlled English'. Paper presented at the Fourth Controlled Language Application Workshop (CLAW 2006), Boston Marriott, Cambridge, MA, USA. 11 pp.

Society of Automotive Engineering (SAE). 2001. *Surface Vehicle Recommended Practice.* Available at: www.lisa.org/useful/2001/J2450Practice.pdf. [Last accessed: 18/11/2004]. 11 pages.

Somers, H. 2003. 'Machine Translation: Latest Developments'. In R. Mitkov (ed.) *The Oxford Handbook of Computational Linguistics.* New York: Oxford University Press. pp. 512-528.

Spyridakis, J. 2000. 'Guidelines for Authoring Comprehensible Web Pages and Evaluating Their Success'. In *Technical Communication,* Volume 47 (3). pp. 301-310.

Spyridakis, J.H., H. Holmback, and S.K Shubert. 1997. 'Measuring the translatability of Simplified English in procedural documents'. In *Transactions on Professional Communication,* IEEE Volume 40, Issue 1, March 1997. pp. 4 - 12

Spyridakis, J., C. Wei, J. Barrick, and E. Cuddihy. 2005. 'Internet-Based Research: Providing a Foundation for Web-Design Guidelines'. In *IEEE Transactions on Professional Communication,* Volume 48, No. 3, September 2005. pp. 242-260.

Sun Microsystems. 2003. *Read ME First!: A Style Guide for the Computer Industry.* Mountain View: Sun Technical Publications.

Systran, 2005. *SYSTRAN 5.0 User's Guide.* Paris: Systran.

Torrejón, E. and C. Rico, 2002.'A New Teaching Scenario Tailor-made for the Translation Industry'. In *Proceedings of the BCS-EAMT Workshop: Teaching Machine Translation,* UMIST, Manchester, UK, November 14-15, 2002 .

Tucker, A.B. 1987. 'Current Strategies in MT research and development'. In S. Nirenburg (ed.) *Machine Translation: Theoretical and Methodological issues.* Cambridge: CUP. pp. 22-41.

Turian, J., L. Shen, and D. Melamed. 2003. 'Evaluation of Machine Translation and its Evaluation'. In *Proceedings of MT Summit IX,* New Orleans, USA. 8 pp.

Underwood, N.L. and B. Jongejan. 2001. 'Translatability Checker: A Tool to Help Decide Whether to Use MT'. In *Proceedings of MT Summit VIII,* Santiago de Compostela, Spain, 18-22 September 2001. pp. 363-368.

van der Eijck, P., M. De Konig, and G. Van der Steen. 1996. 'Controlled Language Correction and Translation'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 64-73.

Van der Meer, J. 2006. 'The Emergence of FAUT: Fully Automatic Useful Translation'. Keynote Address at the 11th Annual Conference of the European Association for Machine Translation, Oslo University, Norway.

Van der Meer, J. 2003. 'At Last Translation Automation Becomes a Reality: An Anthology of the Translation Market'. *Proceedings of EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 180-184.

Van Slype, G. 1979. 'Critical Study of Methods for Evaluating the Quality of MT. Technical Report BR19142, Bureau Marcel van Dijk/European Commission (DG XIII), Brussels. Available at: http://www.issco.unige.ch/projects/isle/van-slype.pdf. [Last accessed: 23/02/2006].

Vassiliou, M., S. Markantonatou, Y. Maistro, and V. Karkaletsis. 2003. 'Evaluating Specifications for Controlled Greek'. In *Proceedings of EAMT-CLAW-03*, Dublin City University, Dublin, Ireland, 15-17 May 2003. pp. 185-191.

Warburton, K. 2003. 'Terminology for Machine Translation'. Presentation given at the LISA Forum, Washington DC, 2003.

Warburton, K. 2001. 'Globalization and Terminology Management'. In S.E. Wright and G. Budhin (eds.) *Handbook of Terminology Management,* Vol.2. pp. 677-696. Amsterdam: John Benjamins Publishing.

W3C, 1999. 'Web Content Accessibility Guidelines 1.0'. Available from: http://www.w3.org/TR/WCAG10. [Last accessed:18/08/2005].

White, J. 2003. 'How to Evaluate Machine Translation'. H. Somers (ed.) *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company. pp. 211-244.

White, J., T. O'Connell, and F. O'Mara. 1994. 'The ARPA MT Evaluation Methodologies: Evolution, Lessons/ and Future Approaches'. In *Technology Partnerships for Crossing the Language Barrier, Proceedings of the First Conference of the Association for Machine Translation in the Americas,* Columbia, Maryland. pp. 193-205.

Wojcik, R.H., J.E. Hoard, and K. Holzhauser. 1990. 'The Boeing Simplified English Checker'. In *Proceedings of the International Conference, Human Machine Interaction and Artificial Intelligence in Aeronautics and Space '.* Toulouse, France. pp. 43-57.

Wojcik, R. and H. Holmback. 1996. 'Getting a Controlled Language Off the Ground at Boeing'. In *Proceedings of the First Controlled Language Application Workshop (CLAW 1996),* Centre for Computational Linguistics, Leuven, Belgium. pp. 114-123.

Wojcik, R., H. Holmback, and J. Hoard. 1998. 'Boeing Technical English: An Extension of AECMA SE beyond the Aircraft Maintenance Domain'. In

Proceedings of the Second Controlled Language Application Workshop (CLAW 1998). Pittsburgh, Pennsylvania. pp. 114-123.

Yang, J. and E. Lange. 2003. 'Going live on the Internet'. In H. Somers (ed.) *Computers and Translation: A Translator's Guide.* Amsterdam: John Benjamins Publishing Company. pp. 191-210.

Yunker, J. 2003. *Beyond Borders: Web Globalization Strategies.* Indianapolis: New Riders Publishing.