Université d'Ottawa · University of Ottawa

Université d'Ottawa · University of Ottawa

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Cathryn Anne ARNOLD

AUTEUR DE LA THÈSE - AUTHOR OF THESIS

M.A. (Translation)

GRADE - DEGREE

School of Translation and Interpretation

FACULTÉ, ÉCOLE, DÉPARTEMENT - FACULTY, SCHOOL, DEPARTMENT

TITRE DE LA THÈSE - TITLE OF THE THESIS

Towards an Evaluation Methodology for Machine Translation Output

L. Bowker

DIRECTEUR DE LA THÈSE - THESIS SUPERVISOR

CO-DIRECTEUR DE LA THÈSE - THESIS CO-SUPERVISOR

EXAMINATEURS DE LA THÈSE - THESIS EXAMINERS

J. Quirion

R. Roberts

J.-M. De Koninck, Ph.D.

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

DEAN OF THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Towards an Evaluation Methodology for Machine Translation Output

by
Cathryn Arnold

School of Translation and Interpretation
University of Ottawa

Under the supervision of
Lynne Bowker, PhD
School of Translation and Interpretation

Thesis submitted to
the Faculty of Graduate and Postdoctoral Studies
of the University of Ottawa
in partial fulfillment of
the requirements for the degree of MA (Translation)

Your file Votre référence
ISBN: 0-494-01400-8
Our file Notre référence
ISBN: 0-494-01400-8

Acknowledgments

A thesis is by nature a collaborative effort. Accordingly, there are many people and organizations to whom I owe my thanks and a debt of gratitude.

First, I would like to thank the School of Translation and Interpretation and the Faculty of Arts at the University of Ottawa for generously providing me with financial support. I would also like to thank the professors and support staff at the STI for their help and assistance.

I owe a large thank-you to my family and friends who provided moral support during the ups and downs over the last two years.

I would also like to thank my four "testers," Mavis Cavanaugh, Melissa Ehgoetz, Kelli Fraser and Francie Gow, who all readily (and cheerfully) agreed to test the methodology and did a fantastic job. Thanks also to Karine Circé for translating my abstract.

And last but certainly not least, I owe a large debt of gratitude to my thesis supervisor, Professor Lynne Bowker, without whom I would never have finished this project. Thank you for your insight, feedback, support and, above all, your patience, during the last two years.

Cathryn
April 2004

Abstract

The exponential increase in communications in many different languages brought about by globalization has resulted in a corresponding demand for translation to be done more quickly, but without sacrificing quality. However, there are currently not enough qualified human translators to keep up with the demand. One way in which translators are trying to cope is by turning to technology, including machine translation (MT) systems. MT has the advantage of being able to produce a large volume of translation in a very short time, but it does not always produce high-quality translations. For this reason, MT cannot replace human translation, but it can make translators' work easier by producing rough drafts. Since there are currently many MT systems on the market, there is a real need for an evaluation methodology for translators to help them choose a system that will best meet their needs. As yet, no universally accepted evaluation methodology exists. The objective of this thesis is to develop and test an evaluation methodology that average translators can use to compare off-the-shelf MT systems and select the appropriate one.

A preliminary design for the evaluation methodology is developed, which includes metrics for evaluating the quality of the raw output produced by any number of MT systems according to the criteria intelligibility, coherence and accuracy. Scales are developed which allow the translator to assign scores to the raw MT output for each of the three criteria. These scores are then accumulated and normalized. In a further step, the raw MT output is post-edited by the translator, who keeps track of the time required to turn this output into a client-ready translation. When taken together, the overall quality-related scores and the post-editing time required provide translator with information that he or she can use to declare which MT system is better for his or her purposes.

With regard to testing, the preliminary methodology is first tested in a pilot study. The data is analyzed and interpreted, with refinements being made based on the results and observations. The refined methodology is applied in a larger-scale test in order to verify and validate the results of the pilot study. Following the larger-scale test, overall conclusions are drawn, which include making additional refinements and suggestions for future research. The general conclusion is that the evaluation methodology proposed in this thesis can indeed be useful for helping a translator to select an appropriate MT system.

Résumé

La mondialisation a entraîné une augmentation exponentielle des communications dans de nombreuses langues, ce qui a fait en sorte de créer une demande tout aussi importante pour des services de traduction plus rapides, mais de qualité. Actuellement, il n'y a toutefois pas assez de traducteurs compétents pour répondre à la demande. Pour les traducteurs, une des façons de composer avec cette demande est d'avoir recours à la technologie, y compris les systèmes de traduction automatique (TA). La TA a l'avantage de pouvoir traduire une grande quantité de documents en très peu de temps; elle ne produit cependant pas toujours des traductions de haute qualité. La TA ne peut donc pas remplacer les traducteurs, mais elle peut faciliter leur travail en produisant des brouillons. Puisqu'il y a de nombreux systèmes de TA sur le marché, il est vraiment nécessaire d'élaborer une méthode d'évaluation pour aider les traducteurs à choisir le système qui répondra le mieux à leurs besoins. À ce jour, il n'existe toujours pas une méthode d'évaluation universellement acceptée. La présente thèse a pour objet d'élaborer et de

mettre à l'essai une méthode d'évaluation dont n'importe quel traducteur pourra se servir pour comparer des systèmes commerciaux de TA et choisir celui qui lui convient.

Une méthode d'évaluation préliminaire est élaborée, laquelle comprend des paramètres pour évaluer la qualité des traductions brutes produites par n'importe quel système de TA en fonction des critères suivants : intelligibilité, cohérence et justesse. Des échelles sont élaborées pour permettre au traducteur d'attribuer une note aux traductions brutes produites par les systèmes de TA selon chacun des trois critères. Ces notes sont par la suite additionnées et pondérées. Ensuite, les traductions brutes produites par les systèmes de TA sont postéditées par le traducteur, qui calcule le temps requis pour rendre ces traductions livrables. Ensemble, les notes globales relatives à la qualité et le temps de postédition donnent au traducteur l'information dont il a besoin pour décider quel système de TA lui convient le mieux compte tenu des ses besoins.

Pour ce qui est de la mise à l'essai, la méthode préliminaire est d'abord mise à l'essai dans le cadre d'une étude pilote. Les données sont analysées et interprétées, puis des améliorations sont apportées en tenant compte des résultats et des observations. La méthode améliorée est appliquée dans le cadre d'une mise à l'essai de plus grande envergure afin de vérifier et de valider les résultats de l'étude pilote. Après cette mise à l'essai, des conclusions générales sont tirées, et des améliorations supplémentaires et des suggestions sont proposées en vue de la recherche qui pourra être faite à l'avenir. La conclusion générale est que la méthode d'évaluation proposée dans la présente thèse peut effectivement aider les traducteurs à choisir un système de TA qui leur convient.

Table of Contents

Chapter 0 Introduction

One of the major impacts of globalization has been an exponential increase in communications in many different languages. As pointed out by van der Meer (2003: 181-182), more and more companies are trying to sell their products to customers abroad, which means that Web sites, marketing materials, product documentation, and sometimes even the products themselves (e.g. software) must be translated into other languages. In addition, because the products cannot be sold to foreign markets until the accompanying translation work has been completed, companies are pressuring translators to work more quickly (while still maintaining high quality, of course) in order to facilitate these foreign sales and thus increase profits. As a result, the demand for translation, and fast translation, has also increased exponentially. There are currently not enough qualified human translators to keep up with the sheer volume of translation that needs to be done (Andrés Lange and Bennett 2000: 203).

One way in which translators are trying to meet the demand for higher volumes and quicker turnaround times is by turning to technology for help. The recently published *Survey of the Canadian Translation Industry* (1999: 39) predicts that translation technology is the sector of the translation industry that will experience the strongest growth, nearing 50% per year.

Translation technology is generally broken down into two main strands: computer-aided translation (CAT) and machine translation (MT). CAT tools can take several different forms, ranging from word processors and electronic dictionaries, to corpus analysis tools and translation memories. CAT tools aim to support translators in their task, either by helping them to improve the quality of their work (e.g. terminology

management systems can be used to facilitate consistency) or by helping them to increase their productivity (e.g. translation memory systems help translators to work faster by allowing them to "recycle" previously translated material). In contrast, MT systems attempt to actually produce translations without any human intervention. The tool that will be the focus of this thesis is the machine translation system[1], which according to the *Survey of the Canadian Translation Industry* (1999: 39) accounts for almost 50% of the translation technology sector.

0.1 Background information and motivation for research

One significant advantage offered by MT is that it can produce a large volume of translation in a very short time. However, a criticism that is commonly levelled at MT technology is that it does not produce "perfect" or even high-quality translations. For this reason, MT cannot replace human translation (HT). However, it can make a translator's work easier and more efficient by eliminating tedious tasks such as dictionary look-up/lexical queries and ensuring terminological consistency (Schäfer 2003). MT can also make translating a text a lot easier because the MT output can be used as a rough draft of the translation. Human assistance may be required after the translation stage, typically through a process known as post-editing, which is carried out in order to "clean up" the MT output.[2]

---

[1] Although further discussion of CAT tools is beyond the scope of this thesis, readers can refer to sources such as L'Homme (1999), Austermühl (2001) and Bowker (2002) for additional information.

[2] Pre-editing is another approach to producing high-quality machine translation. Instead of "cleaning up" raw MT output, texts to be translated are revised by a human editor who uses a controlled or restricted vocabulary and syntax as a pre-emptive measure intended to reduce ambiguity before the MT system even begins to process the text, resulting in better quality output. See Lockwood (2000) and EAMT-CLAW (2003) for more information on pre-editing.

Since MT typically produces less-than-perfect output, some critics question whether it is suitable for processing all types of texts; however, for some types of texts, MT has proved to be quite adequate. While MT is unlikely to be used for translating creative texts (e.g. literature, advertisements), it can be useful for producing an initial draft of more informative-style texts such as administrative, technical or specialized texts. For instance, Brace (2000: 219-220) reports that the European Commission, which is the administrative wing of the European Union, has a translation department that produces over 1 million pages a year, and in 1996, some 220,000 pages were run through the Systran MT system. When translation is required simply for information purposes, the raw MT output may suffice; however, if the target text needs to be more polished (i.e. for dissemination to a wider audience), then a human translator may need to step in and do some post-editing. If the amount of time required to post-edit the text is less than the time that would be required to translate it from scratch, then a time saving has been realized and the translator can use the time saved to work on another job. In such a case, technology can help a translator to meet the dual demands for translating more quickly and for translating a greater volume of text. One company that has made MT work successfully for them is the Baan Company in Germany, who report that, if everything runs smoothly, the time required to produce a translation can be cut by 50-60% with the help of an MT system (Andrés Lange and Bennett 2000: 208).

Even if they accept that MT can in fact be useful in some translation situations, translators who are interested in using this technology are still faced with a difficult choice: which system is "best"? This is a valid question since the translator will have to invest money in purchasing the system and time in learning how to use it. However, there

is no easy answer since translation contexts can vary widely (e.g. different subject fields, text types, language combinations) and different tools provide different features or approaches that may work well in some contexts and less well in others. Therefore, what is "best" for one translator may not be best for another and it is important for a translator who is considering purchasing an MT system to carefully evaluate its production or output in the context of his or her own needs rather than simply relying on a recommendation from a colleague or a software review, etc. Since there are currently many MT systems commercially available[3], and since, as pointed out by Arnold et al. (1994: 180) "The purchase of an MT system is in many cases a costly affair and requires careful consideration", there is a real need for an evaluation methodology that translators can use to help them select an appropriate one. Furthermore, this evaluation process should be reusable/generic and comparable so that reliable and realistic evaluations can be made of any MT system in order for potential purchasers to be able to make the optimal decision as to which of the many systems will best meet their needs.

A preliminary survey of the literature in the field of MT evaluation reveals that there is no standard or universally accepted rating system. Although a lot of work has been done on evaluation metrics and there is general agreement as to which criteria can and should be measured, no reusable evaluation methodology has thus far been developed. Additionally, the growing use of MT systems has only increased the need for comparable evaluation metrics (White 2003: 211). Although no evaluation methodology can definitively answer the question "Which tool is best?", a framework that allows

[3] The list of commercially available MT systems that are currently on the market includes Systran, Logos, Reverso Pro, LogoMedia and L&H Power Translator Pro among others.

potential users to evaluate MT systems according to specific criteria could at least help them to choose a tool that meets their most immediate needs.

There are various groups of stakeholders interested in MT software systems, each of which is concerned with different aspects. These groups include investors, software researchers and developers, purchasers, translators and their clients or end-users. Each of these groups has its own goals and operates under specific conditions and constraints so that their evaluation needs are different (Trujillo 1999: 251-256). For example, researchers and developers are more interested in the development of MT systems. While researchers need an evaluation methodology that will enable them to build and debug a prototype, investors are concerned with the feasibility and economic potential of the MT system under development.

On the other hand, purchasers, translators and their clients are more concerned with the practical or functional aspects of MT systems. For example, purchasers, who as a group range from individual translators to large translation departments or companies, are more interested in the quality of the MT system translations (output), its speed, efficiency and degree of user-friendliness, ease of updating and customizing its dictionaries and how well it can be integrated with other software packages. Translators' clients, like purchasers, are concerned with a system's cost, speed and quality of its output, but are not particularly concerned with whether or not its internal dictionaries can be updated or if it is compatible with other software packages.

Thus, there are many aspects or features of MT software that can and need to be evaluated depending on the stakeholder. Different evaluation strategies are used

depending on who is doing the evaluation and what is being evaluated. Generally evaluation strategies can be divided into two main types: glass-box and black-box. If the evaluator has access to the internal workings, the underlying algorithms and source code, the evaluation strategy is glass-box. Glass-box evaluation is used to look at the various components of an MT system and how they affect the overall workings of system. Generally, researchers and developers will need to use a glass-box evaluation strategy, since they need access to the source code to fix or improve the system.

Conversely, black-box evaluation looks at MT operations uniquely in terms of input (ST) and output (TT) (Trujillo 1999: 256). The inner workings, underlying algorithms or actual source code are not accessible to the evaluator. This strategy is relevant for users and translators, who are not in a position to directly improve the software or fix any of its faults. Black-box evaluation is used to assess the functionality of an MT system, the volume of translation produced, quality of the output and user-friendliness, all of which are extremely important to the end-user of a system.

0.2 Objectives

The objectives of this thesis are to *develop* and *test* an evaluation methodology for assessing the quality of MT output.

The evaluation methodology will be generic, comparative, task-based and declarative. These characteristics can be defined as follows:

- **Generic:** A generic or reusable system refers to the case where a single evaluation methodology can be used to evaluate the output of any MT system regardless of its architecture, the language-pair or the text type. Using the same—

or generic—methodology for all evaluations will ensure that the results are comparative.

- **Comparative:** In the context of this thesis, a comparative evaluation is one that compares the raw translation of the same source text done by multiple MT systems. It does *not* refer to a case where MT output is compared with translation(s) done by human translators (HT output).
- **Task-based:** In addition, the evaluation methodology developed in this thesis will be task-based in that a specific user group and the purpose for which it requires the translation will be selected as the basis for evaluation (EAGLES 1999, 1996).
- **Declarative:** A declarative evaluation methodology is one that will measure the ability of MT systems to handle "real" texts and allow the translator-user to declare which system seems "better" for the task at hand. According to White (2003: 219), declarative evaluation measures the intelligibility and fidelity of translations.

0.3 Scope and limitations

As established in section 0.1 above, there are a number of different stakeholders interested in MT evaluation, as well as a number of different features that can be evaluated, which may require different approaches to conducting evaluations. It is not feasible to address all these issues in a single MA thesis, so I have chosen to place a number of limitations on my research.

0.3.1 General evaluation approach

Firstly, the type of MT evaluation methodology that will be developed as part of this thesis will be of the black-box variety (see section 0.1), and its scope will be limited to analyzing the quality of the raw MT output or translation produced by an MT system.

0.3.2 Task scenario

In addition, as noted in section 0.2, one of the characteristics of the evaluation methodology to be developed is that it will be task-based. This means that a specific user group with a relatively circumscribed task (i.e. specific text type and function) must be identified.

Since the mid-1990s, research into developing an evaluation method for MT has been focusing on a task-based approach (i.e. defining the parameters and criteria of the testing) as opposed to trying to develop a "one-size fits all" metric that can be applied no matter what the circumstances or the task. A task-based approach is efficient in that it will narrow the scope of the preliminary testing, with the result that it can be carried out more quickly. If the results are promising, the evaluation methodology can then be applied to a wider range and larger corpus of texts. Task-based approaches have been more successful in general, as the "one size fits all" approach has proven to be too vast and complicated for MT evaluation as the results are so general as to be almost useless (Vanni and Reeder 2000: 109). Furthermore, a task-based evaluation is reasonably realistic since the majority of translators tend to specialize and so deal with a relatively limited number of text types and subject areas on a regular basis, which means that it is

possible for them to come up with reasonable parameters and criteria to evaluate translated text. I have chosen to constrain the research carried out in this thesis by restricting the development and testing of the evaluation methodology to the following "task scenario," which I will describe in terms of the intended user group, the text type, and the systems and language pairs to be tested.

0.3.2.1 User group

Overall, my aim is to develop and test the metrics based on the most realistic conditions possible. Since I have been trained as a translator and do some freelance translating, I decided to look at the issues involved in MT system output from the perspective of a translator-user working either on a freelance basis or for a small translation company in Canada. According to the *Canadian Survey of the Translation Industry* (1999: 11), there are approximately 6,040 translators who work either on a freelance basis or for a small firm, as compared to 6,490 translators working for large firms (including the Government of Canada's Translation Bureau). Though the number is divided almost evenly, I feel a comparative evaluation metric that is quick and easy to perform would prove to be particularly useful to the freelance group of users.

On the whole, this type of translator tends to specialize in specific types of translation (including specialized translations). As well such translators tend to translate a large volume (i.e. more than 2,000 words per day). They are also the ones that would greatly benefit from using an MT system as this would allow them to increase their production. More production means more revenue for them. While it is true that using an MT system could also benefit in-house translators at a large company in the sense that

they could also produce more translations, they may not benefit directly to the same extent as would freelance translators. Furthermore, freelance translators and those working for small companies do not have a lot of resources—that is, either the time or money—to invest in purchasing and learning how to use a range of MT systems. Therefore they would also benefit from a comparative MT evaluation methodology as it would allow them to make the optimal choice for themselves or their companies.

0.3.2.2 Texts

I decided to develop and test my evaluation methodology using texts that freelance translators working in Canada would realistically be asked to translate: specialized, informative or administrative texts. I will also use a variety of document types because normally, a translator would be required to translate more than one type of document. Furthermore, I will use entire documents as opposed to test suites[4] since translators do not translate sentences out of context or grammatical problems—they translate entire documents. Also, test suites are more suitable for glass-box evaluations where evaluators need to know the inner workings of the MT system so that they may concentrate on a particular part of the system. Full texts are better for black-box evaluations where the emphasis is on the functionality of the MT system and its output.

[4] According to Trujillo (1999: 257-258), test-suites involve the use of a carefully constructed set of examples (sentences) designed to test specific linguistic phenomena or translation difficulties. Each sentence is context independent. The major advantage of using test-suites is that they test known translation problems, which may or may not be present in all texts and so provide better coverage of known problem areas. However, actual translations of actual documents in context are not examined. Test-suite evaluation systems are based on the assumption that the "behaviour" of an MT system can be projected from results of test suite evaluations. Therefore, they only indirectly evaluate output of MT systems. Furthermore, the results of test-suite evaluations are difficult to compare because one system may handle a specific problem better than another, but not be as effective for another problem area. This raises the issue of which type of problem is more significant.

The specific set of texts used to design and test this methodology will be described in more detail in sections 3.2 (pilot test) and 4.1 (larger-scale testing). Sample texts are included in Appendix A.

0.3.2.3 Systems

Another way in which I will limit the scope of this research project is by imposing selection criteria and limiting the number of MT systems to be evaluated. Ideally, the evaluation methodology developed in this project will be generic enough to allow users to compare the output of any available MT systems; however, for practical reasons of time and system availability, the development and testing conducted as part of this thesis will be limited to using the following two systems: Systran Classic (1999 version) and ReversoPro 98. The following criteria were applied in the selection of these two MT systems:

- available commercially. In other words, be readily accessible to translators.
- affordable to a freelancer or small translation agency.
- have modest system requirements (i.e. run on a standard desktop PC with a Pentium processor and 256K of RAM).
- compatible with one or more popular word-processing packages to facilitate post-editing.
- able to handle French-to-English language combination.
- allow users to update dictionaries.

In addition, these two systems were available to me at the STI, and although the versions used are not the most recent versions of the systems in question, this does not matter for

my purposes since the aim of this thesis is not to actually recommend the best MT system for freelance translators to use, rather, the aim is to develop an evaluation methodology that would allow translators to reach their own conclusions about the suitability of any systems they choose to evaluate.

0.3.2.4 Language pair

Finally, although the idea of this thesis is to devise an evaluation methodology that could, in principle, be applied to any language pair, I will limit my investigation to French-to-English translation because I am a French-English translator and this is the language direction that I am competent to research and work in. Furthermore, it is one of the most common language combinations used in the Canadian translation industry.

0.4 Methodological approach

The general methodological approach employed in the context of this thesis can be divided into a number of steps as follows:

1. literature survey to determine what researchers have already learned about evaluating MT output.
2. design of initial evaluation methodology, including definition of top-level characteristics and metrics. These will be based upon a combination of information gleaned from the literature survey and practical experimentation.

3. pilot test of the initial evaluation methodology. A text will be run through the two MT systems and their output will be scored using the metrics developed as part of step 2 above.
4. assessment of pilot test results and refinement of the evaluation methodology.
5. testing of the refined evaluation methodology. A wider variety of texts will be run through the two MT systems and their output will be scored by independent evaluators, but this time using the refined evaluation methodology.
6. analysis of results and assessment of refined evaluation methodology.
7. suggestions for future research.

0.5 Outline

Chapter 1 provides an overview of translation evaluation in general as well as issues and difficulties specific to the evaluation of MT output. Chapter 2 outlines the preliminary design for the proposed evaluation methodology including a discussion on the objectives and the metrics to be used in the evaluation. The results of a pilot test and refinements to the proposed methodology are described in Chapter 3. In Chapter 4, the refined methodology is applied in a larger-scale test whose purpose is to validate and verify the results of the pilot study. Additional refinements to the methodology are also proposed in Chapter 4. Chapter 5 reports on the conclusions and suggestions for possible areas of future research. The main body of the thesis is followed by a glossary, a list of references and a number of appendices containing the materials used in and resulting from the tests.

Chapter 1 Theoretical Framework

1.1 General overview of evaluation of translation

Before starting to examine the difficulties and issues specifically related to evaluating the quality of MT output, I will look at the difficulties and issues related to evaluating translation in general. Because a translated document is a translation regardless of whether it has been produced by a human or a computer, the problems inherent in evaluating human translation (HT) also apply to MT. In fact, according to Trujillo (1999: 258), measuring the quality of translation is not only difficult for MT, but also for other types of translation.

1.1.1 Difficulties and issues

The difficulties and issues regarding translation and translation evaluation stem from the fact that for any ST, there are potentially many good translations; that is, there is no one "correct" translation. The fact that there are many diverse theories of translation compounds the situation. Each particular theory will give rise to many potential different translations. Not only that, translators who subscribe to one theory will not necessarily find translations done by others subscribing to a different theory to be good quality or even acceptable. And since there is no consensus as to what makes a good translation, it seems only logical that there are no standards for evaluation or for a method of evaluation.

1.1.1.1 No universally accepted theory of translation

As discussed in section 1.1.1, one of the major problems in evaluating translation arises from that fact that for any given document or text, there is no one "perfect" or "correct" translation. According to Schäffner (1998a: 1) the concept that the aim of each translation activity (i.e. translating a text) is to produce a good translation or good target text presupposes that there are criteria for deciding just what characteristics result in a good translation. However, no specific mention is made of what these criteria are that enable someone to say that one text is a "good" translation as compared to another. Accordingly, there is no absolute or "gold standard" that a translation (whether done by a human or a machine) can be measured against, although translations can be rated as being generally good or bad.

Indeed criteria for measuring the quality of a translation vary depending on the purpose of the assessment and the theoretical framework applied to the translation. As House (1997: 1) states, "Evaluating the quality of a translation presupposes a theory of translation." Accordingly, there are as many different concepts of quality and, by extension, criteria for evaluating it as there are theories or views of translation quality.

The fact that there are many different schools of thought about or approaches to translation, each one placing varying degrees of emphasis or importance on different aspects of translation, including translated texts, complicates the issue of translation evaluation. Savory has compiled a list of what various translators and translation scholars believe are the qualities of a good translation, which illustrates this point:[5]

1. A translation must give the words of the original.

[5] From T. H. Savory (1968) *The Art of Translation,* Cape, 54, cited in Newmark (1988).

2. A translation must give the ideas of the original.
3. A translation should read like an original work.
4. A translation should read like a translation.
5. A translation should reflect the style of the original.
6. A translation should possess the style of the translation.
7. A translation should read as a contemporary of the original.
8. A translation should read as a contemporary of the translation.
9. A translation may add to or omit from the original.
10. A translation may never add to or omit from the original.
11. A translation of verse should be in prose.
12. A translation of verse should be in verse.

As can be seen from this list, there is no overall agreement on what it takes to be a "good" translation. Some translation scholars feel that the way a translated text reads is of the utmost importance, while others feel that it should be fidelity to the original. In other words, some evaluators feel style is very important while others feel that if the translation conveys all the information contained in the original with no meaning errors, then the translation is a good one.

Yet, according to Sager (1994: 142) the concept of equivalence is a fundamental element in all theories of translation. What varies is how this concept has been interpreted. It is generally agreed that there are three types of equivalence (and relationships) between an ST and TT: cognitive, linguistic and pragmatic. Ideally, all three types should be achieved at text level. The problems and issues arise in deciding on

how these three types of equivalence should be achieved, which has resulted in the development of various approaches to translation.

Approaches to translation, and by extension approaches to evaluating translation quality, can be grouped into four main categories as follows:[6]

1. Anecdotal, biographical and neo-hermeneutic approaches.

Most of these approaches to evaluation are based on the concept that the quality of a translation depends largely on the translator's interpretation and transfer decisions, which are based on his/her linguistic and cultural knowledge and skills. Accordingly these approaches are subjective and vary from translator to translator. (House 1997: 3)

2. Response-oriented and behavioural approaches.

Adherents of these approaches do not agree with concept that it is the translator's creative skills that will result in quality translations. As reported in House (1997: 4), Nida is a proponent of this type of approach and has proposed the following three criteria for assessing translation quality:

-general efficiency of the communication process,

-comprehension of content, and

-equivalence of response, which is closely related with his concept of dynamic or functional equivalence.

3. Linguistic approaches.

Proponents of linguistic theories of translation believe that the ST (i.e. its meaning or informational content) plays a predominant role in the translation process. Linguistic approaches include many different schools: some are solely linguistic, while others encompass pragmatic, socio-cultural and discourse aspects of the ST as well (House

[6] For a more detailed discussion on the major approaches to translation, see House (1997).

1997: 16). Newmark's concept of semantic translation, in which the most important criterion is the transferring of the contextual meaning of the ST, is an example of a linguistic approach to translation (Newmark 1988: 22).

4. Text-based approaches.[7]

According to House (1997: 6-16), there are three major schools of text-based approaches. The first main category is the literature-oriented approaches that can be adapted for evaluation, including descriptive and polysystem theory. Toury is a proponent of polysystem theory. The second category includes the post-modernist and deconstructionist approaches (proponents include Derrida and Venuti), in which the goal is to rethink translation from a philosophical and sociological point of view. This approach includes what House calls overt translation, that is, where a translation is obviously a translation and the ST becomes less important. The third school includes the functionalist[8] and action and reception theory approaches (e.g. Skopos theory). Functionalists include Reiss, Vermeer and Hönig. Adherents of these theories believe that the purpose or function of the translation is the most important factor and dictates the translation strategy adopted. Again, in this category, the ST loses some of its importance.

Underlying all these approaches to translation (and by extension, to translation quality evaluation) is the fundamental dichotomy between literal and free translation, which has existed since translation began. According to Melby and Warner (1995: 8-11) and Horguelin and Brunette (1998: 11-13), over the past 2,000 years, translation theories have been concerned mainly with the tension between literal translation and free

[7] For a more detailed discussion on text-based and pragmatic approaches to translation theory, see Schäffner (1998b).
[8] For a more detailed discussion on functionalist approaches to translation theory, see Schäffner (1998b) and Hönig (1998).

translation, which represent two extremes on a continuum. It is rare to see a translation that is either totally free or totally literal. Most translations lie somewhere along the middle of the continuum (i.e. they are a mixture of both free and literal translation).

Indeed, Fernández Guerra (2000: 21) posits that no matter what position a translator may take on the free/literal dichotomy, he/she is forced to compromise because:

1. complete fidelity to the source text is not possible due to the fact that meaning and designation may not coincide in the TL. For example, the phrase *two heads are better than one* may be translated into French literally (i.e. its designation is translated) but the meaning is not the same at all; and

2. complete functional equivalence to the source text is not possible either because the referent may not actually exist in the TL culture (e.g. a Chinese "novel of manners" is a format/genre that does not exist in English). She adds that there are three "golden" rules of translation: 1) don't omit anything; 2) don't add anything; and 3) transfer the content in such a way that readers will not be aware they are reading a translation. The problem arises when translators try to follow all three rules at the same time. Since this is clearly impossible, translators must compromise.

Newmark (1988: 113) states that there is no such thing as a law of translation, since laws admit no exceptions. There can be no valid single comprehensive theory of translation and no general agreement on the element of invariance, the ideal translation unit, the degree of translatability of a text and the concepts of equivalent-effect and congruence in translation. However it is worth striving to achieve all these concepts.

Accordingly, Newmark feels that there is no such thing as a science of translation, that is, translation is not an exact science and never will be.

1.1.1.2 Distinction between criticism, revision and evaluation

On an even more fundamental level lies the issue of exactly what is evaluation. A distinction can be made between criticism, revision and evaluation (Horguelin and Brunette 1998: 3).

According to Horguelin and Brunette (1998: 5-6) criticism applies to literary and artistic texts that are already published or finished, and as such are "evaluated" after the fact. House's model for evaluating translation[9] quality belongs to the category of criticism (House 2001: 243-257).

Revision applies to texts before they are considered to be finished products and can be defined as the process of controlling document production for accuracy, completeness, stylistic appropriateness, etc. It deals with form at various levels such as spelling, punctuation, paragraphs, and document layout. According to Sager (1994: 238), revision is an important aspect of translation and often represents one-third of the total translation effort. Post-editing is the most recent form of revision and is specific to MT (Horguelin and Brunette 1998: 5). The nature of the revision depends on the quality of output required.

As is the case with criticism, evaluation applies to finished texts. It should also be noted that not all translations are evaluated. While most human translations are revised (Horguelin and Brunette 1998: 9-10), it appears that they are not all evaluated. According to Martínez Melis and Hurtado Albir (2001: 273), translation evaluation is relevant in

[9] See section 1.1.1.4 for more details on House's model.

three areas: published texts, texts produced by professional translators and texts produced in educational contexts. Evaluation of published texts seems to correspond to what Horguelin and Brunette call criticism. Various methods of evaluating texts produced in educational contexts have been devised.[10]

For the purposes of this thesis, criticism is not relevant as it deals only with published documents. Evaluation is relevant in that the methodology developed will be used to evaluate the quality of the raw output (that is translations in an intermediate, non client-ready, form) in terms of its usefulness as a draft for professional translators. Revision (or in the case of MT, post-editing) is also relevant for the purposes of this thesis in that the post-editing effort required to transform translations into client-ready form will be used as an evaluation metric.

1.1.1.3 No universally accepted standards for evaluation or revision

Given that there is no globally accepted approach to translation and no one correct translation of a text, it seems reasonable that there are no universally accepted standards for translation quality evaluation or for revision. Hönig (1998: 28-29) has looked at a range of evaluation contexts, including language acquisition, translation courses, translator testing, quality assurance in a translation department of a company, and acceptability by end users, etc. A summary of his results indicates that:

1. there are no common criteria for evaluating translation quality;

2. the most common criterion is that of meeting the text production standards in the TL;

[10] See Bowker (2001), Lee-Jahnke (2001) and Waddington (2001) for their approaches to evaluation in educational contexts.

3. in one-half of the scenarios, it is not clear whether the assessment is carried out based on ST or TT orientation;

4. the most homogeneous evaluation criteria are applied in language acquisition and quality assurance (i.e. commercial translation settings) scenarios; and

5. the least homogeneous criteria are applied in university training courses.

Based on these results, it seems that there is no consensus on standards for evaluation or for revision.

The traditional approach to evaluating translation quality has been linguistic, the criteria being that good translations are as accurate as possible (Newmark 1991: 111). However, accuracy is a relational concept: the issue has always been accurate in relation to what. Generally speaking the yardstick has been the ST. According to Newmark (1991: 111), comparing the ST to the TT involves quantitative measures (i.e. measuring the completeness of transfer) and qualitative measures (i.e. measuring the degree of accuracy for denotation and connotation, both referentially and pragmatically).

As noted in section 1.1.1.1, the field of translation studies is not a homogeneous discipline, and the linguistic model has largely been replaced or complemented by other approaches (Schäffner 1998a: 1). Each of these new approaches contributes more insights to the understanding of the translation process and applies different criteria for evaluating translation quality. However, linguistic models have also evolved over the years: according to Schäffner (1998a: 2), House demonstrates that her linguistic approach includes textual, situational and cultural aspects.

Text typological considerations are also an important aspect of evaluating translation quality, but they have been largely ignored in traditional linguistic models.[11] An example of a text-based model is Hatim and Mason's model, which includes three text categories or forms: argumentation, exposition and instruction (Sager 1994: 170). The problem with Hatim and Mason's typology is that it is too simplistic: most texts are hybrids (Horguelin and Brunette 1998: 25). Furthermore, text typologies have not produced any objective means of measuring quality—they have only opened up new ways of thinking about and attacking the problem (Horguelin and Brunette 1998: 30). In other words, there is no universally accepted set of criteria or one "true" method for evaluating quality.

1.1.1.4 Subjective versus objective evaluation

As discussed in section 1.1.1.3, there seems to be no consensus on standards for evaluating translation. There is a current movement towards objective evaluation, but the difficulty lies in finding a methodology that is objective. It seems reasonable that complete objectivity is impossible due to the fact that, for the time being at least, evaluation is always done by humans and therefore reflects their individual bias. In the past (i.e. before translation became an industry approximately 50 years ago), it seems that most evaluation was really criticism and tended to be subjective.

According to Horguelin and Brunette (1998: 11), the development of quality criteria (especially for revision) implies judging the quality of the TT and this judgment should be made according to recognized ("reconnu") specific criteria so as to not be

[11] For a more in-depth discussion of text typologies, see House (1997).

completely subjective. They state that this has been problematic because what is considered to be (or not to be) a good text varies from era to era and from text type to text type. As mentioned in 1.1.1.1 the tension between literal and free translation has always existed, and historically one type or the other has dominated depending on the era and the location (Horguelin and Brunette 1998:11-13). Thus it appears that there is no such thing as immutable criteria.

The issue of objectivity in evaluation was discussed at the *Fédération internationale des traducteurs* (FIT) congress in 1959. Although the goal of objectivity in evaluation was felt to be worthwhile, no concrete criteria emerged from the congress (Horguelin and Brunette 1998: 18). Horguelin and Brunette feel that this indicates that, objectively speaking, translation quality does exist and there should be some way of measuring it. Accordingly, new parameters need to be set—especially since the 1970s given that pragmatic (i.e. commercial or informative) translation has become more prevalent than literary translation as a result of increased communication and globalization.

According to Williams (2001: 236), models for evaluating translation quality fall into two main categories: quantitative (e.g. Sical[12]) and non-quantitative (e.g. House 1997). Williams states that quantitative models have major shortcomings because they tend to focus on microtextual analysis and error counts (i.e. the text as a whole or its macrostructure is not taken into account). However, these models have the advantage of being objective in that criteria are established and then measured. Williams states that non-quantitative models also have major shortcomings and are not wholly acceptable either because they do not propose error weighting and quantification for individual texts.

[12] *Système canadien d'appréciation de la qualité linguistique.*

The Sical scale developed by Darbelnet for the Translation Bureau is quantitative, and as such, more objective. However, when this model was implemented, it was found to have too many parameters and criteria (up to 675, depending on the version), and as a result it was too expensive and cumbersome to implement (Horguelin and Brunette 1998: 20).

Communicative models can also be quantitative and objective. Examples include models that use measures such as cloze tests where the reader is asked to fill in missing words, the Dale-Chall formula which measures readability by calculating the ratio of total words to usual words, the Gunning Formula which measures difficulty by calculating the ratio of total words to words of more than three syllables, and the Flesch method which measures readability by calculating the average length of words and also measures interest to humans using number of "personal" words versus number of "neutral" words. These models are all intended to provide a measure of the effectiveness of communication. Yet, according to Horguelin and Brunette (1998: 27-28), these methods are not always accurate because they do not take into account the situation of communication.

Yet, not all research is geared towards developing a methodology that assigns values or quantifies translation (Horguelin and Brunette 1998: 21). Two notable exceptions are the models developed by Nida and Taber (1969: 167-170) and by House (1997).

Nida and Taber (1969) developed the concept of dynamic equivalence as opposed to formal correspondence. According to their model, translations must be intelligible and readable, and must provoke the same reaction in the TT reader as in the ST reader (their

major criteria for a good translation). Good quality translations will contain a high degree of dynamic equivalence. Nida and Taber recommend the use of cloze tests to evaluate translation, but add that more practical means also need to be used as these tests may be difficult to administer and the results cumbersome to analyze. These practical measures include judging receptors' reactions to alternative translation options, having them explain the contents of the translation and having the subject read the translation aloud. They also state that certain statistical measures such as the number of words per sentence and the number of syllables per word can be used, but that such measures are only rough calculations.

The model proposed by House (1997) is pragmatic but has no quantitative criteria either. House's model assesses quality by analyzing the ST, stating its function, analyzing the TT, comparing the TT to the ST and making a statement of quality. The categories for analyzing both the TT and the ST are register (field, tenor and mode) and genre. In order to analyze these categories, House examines the syntactic means, lexical means and textual means (or devices) used to establish equivalence of these categories. Three main textual aspects are examined: theme dynamics (theme-rheme patterns), clausal linkage (logical relations between clauses and sentences –additive, adversative, causal, explanatory, etc.) and iconic linkage (structural parallelism-isomorphic). In addition House's model divides translation into two main types: overt (i.e. obviously a translation) and covert (enjoys status of original text—that is, it reads like an original). The differences between SL culture and TL culture are dealt with by means of a cultural filter in covert translation. It should be noted that House (1997: 118) herself states that her model is not "absolutely evaluative."

Finally, the issue of who is qualified to be an evaluator has implications for objective evaluation, and should therefore be taken into account (Sager 1994: 145-46). For example, Sager states that the ST author and end-users are not qualified because they are unilingual. Neither is the translator of the text in question because he/she is subjective (i.e. each translator has his or her own idea of how a text should be translated).

In summary, as is the case for translation theory, there seems to be no consensus on the approach that should be adopted for evaluating translations. It seems that no simple or universally accepted measures exist for evaluating a translation's quality, and not all approaches are objective. It is beyond the scope of this thesis to provide an answer to the overall problem of translation evaluation. Accordingly, I have chosen to look at quantitative approaches to evaluation of translation quality, which I feel tend to be more objective in the sense that they can be measured or counted, rather than qualitative approaches, which tend to be more subjective. Quantitative approaches measure predefined criteria using guidelines whereas qualitative approaches tend to use vaguely defined criteria to decide on the quality of translations.

1.1.2 General evaluation criteria

As discussed in sections 1.1.1.3 and 1.1.1.4, the development of the translation industry over the past 50 years has resulted in the evolution of approaches to translation and evaluation and has impacted on criteria for evaluating translation quality. According to Horguelin and Brunette (1998: 37-38), historically the criteria for translation quality were a reflection of the time, place and situation of communication, the dominant translation theory in place, type of translation, text type, and the end user. Currently, the

criteria have changed and are exactitude, correctness, readability, functional adaptation and cost effectiveness.

Sager (1994: 148-49) agrees that the criteria are evolving and states that a number of criteria have been advanced for assessing translation in an industrial environment, with the classical criterion being fidelity or accuracy. Sager also states that a number of criteria can be grouped and used to evaluate the acceptability of industrial translations. These criteria include fidelity, intelligibility and usefulness.

According to Trujillo (1999: 258), in general the approach to evaluation is to ask a translator or subject matter expert to rank output texts for quality, both in terms of monolingual intelligibility (i.e. whether or not the text makes sense and whether or not it reads as if it were written by a native speaker) and bilingual accuracy (i.e. whether or not the TT is faithful to the ST and whether or not its meaning has been transferred).

However, the simplest form of evaluation deals with the concept of error, wherein various levels are established (Sager 1994: 240-242). One method includes three levels of gravity (distortion of sense, omission, and minor errors such as stylistic infelicity, spelling, etc.). Other, more elaborate systems include more error classes, which are subdivided into lexical (meaning) or syntactic (language) errors. He feels that these refinements of methods of quality evaluation are too complex to be developed into a practical tool for quality control and are not sufficiently practical because they do not consider time and cost factors that strongly influence the degree of quality. Much simpler methods can be applied when economic considerations are the reason for evaluation. Time spent on re-reading and revision is an indication of both the volume and the gravity of errors.

One of the issues of evaluating fidelity involves the level at which it should be established (i.e. on a linguistic, pragmatic or cognitive level). To ensure fidelity at all three of these levels, the evaluation would have to be carried out using various units, which would include the phrase, clause, sentence, paragraph and text levels. According to Horguelin and Brunette (1998: 20-21), using translation units (TUs) as the base unit to rate translations is not feasible unless translation is solely a linguistic operation wherein the TUs of the ST are reproduced in the TT. Furthermore, they state that using the TU as a base unit entails segmenting the text, with the result that this method tends not to be used due to the prevalence of communication theories, in which the function of the text is important and dictates against the use of TUs. In addition, segmenting a text leads to questions of how to go about it. Meaning does not reside completely in linguistic units: it includes graphic signs and readers' world knowledge. There is also evidence to suggest that translation units vary from individual to individual (i.e. the size of translation units is not an objective concept). According to Sager (1994: 211) since very little is known about the human process of translation, it is generally assumed that translators proceed by somehow identifying cognitive units in the SL and finding equivalent TL units. The process of recognizing and delimiting translation units is assisted by subject knowledge. Units are generally limited by phrase and sentence boundaries in the SL, but can depend on SL text structure.

Indeed the whole issue of fragmenting texts into segments may be an interesting one in the context of MT and CAT because these tools actually *do* segment the text, which may not be the way humans operate. The fact that MT systems generally use the sentence as the translation unit with the result that they do not consider the text as a

whole can be considered an argument in favour of using coherence as a scoring metric for MT system evaluation as discussed in section 1.2.4.

1.2 Issues and difficulties specific to MT evaluation

In addition to the general issues discussed in section 1.1 above, there are other issues and difficulties specific to the evaluation of MT. Using a computer in the translation process fundamentally changes it, and leads to the need for evaluation of aspects of the process *per se*. At the level of quality of output, one of the main reasons difficulties and issues arise is due to the fact that MT systems do not and cannot produce dynamic, living language the way humans do.

As mentioned in 1.1.2, translation evaluation generally consists of measuring output texts for quality, including the two major criteria of intelligibility and accuracy. Although such an approach is applicable for all types of translation (Trujillo 1999: 258), it seems to be more of an issue for MT. If a translation has been done by a human, it appears to be less likely that it will be evaluated for quality. However, MT output is regularly evaluated for quality.

1.2.1 Issue of comparison of MT with HT

According to Arnold (2003: 120), there is an essential dichotomy between accuracy and creativity. Accuracy includes factors such as faithfulness to the ST, transferring content and meaning, etc., whereas creativity includes factors such as form and style. MT systems cannot be creative, but they can and do strive for accuracy and

intelligibility. Sager (1994: 256-258) states that using computers in the translation process introduces a fundamental change in communication, which results in computer-produced texts "written" in a new language, which is an artificial language because it is produced using previously defined rules and lexicons—designed by the programmer(s). The input (or ST) is written in natural language and the output is written in artificial language, which bears strong resemblance to natural language. Texts written in artificial languages must be post-edited to further resemble natural language.

It should be noted that this point of view is not held by all MT researchers. For instance, Hutchins and Somers (1992: 3) define MT systems as "Computerised systems responsible for the production of translations from one natural language to another [natural language], with or without human assistance."

In any case, MT systems can only produce sentences permitted by the rules in their grammars and their lexicons (dictionaries). These rules and dictionaries reflect the language as perceived by the programmer(s)—which is a subset or derivative of natural language. MT systems do not include text grammars (i.e. they work only on a sentence basis), which means there is a potential problem for coherence, which concerns text as a pragmatic whole.

Furthermore, Sager (1994: 258-261) maintains that these artificial computer languages are usually functionally restricted to conveying information and do not convey connotative, emotive, aesthetic or other nuances of language (i.e. they "follow the rules"). They are a useful tool for efficient communication among specialists, which is why MT systems are good for translating informational or specialized texts. Due to the fact that any translation of a natural (source) language text into an artificial (target) language text

(as is the case in MT) implies a reduction in function and pragmatic scope, MT-produced texts are another translation-specific text type. It is currently recognized that MT systems can be optimized for particular text types in specific subject fields and for specific functions and purposes.

Sager (1994: 260-262) argues that the quality of human translation is impacted by the conditions of production including the quality of the ST, the time allowed for translation (the deadline) and the translator's level of skill. Accordingly, there is no single model of human translation that can be used as a standard for MT and there is no minimum standard of HT which can be set as a target for MT. However, the nature and quality of MT was initially defined only in relation to HT and its degree of closeness to HT, which is a fallacious basis for assessment because it is based on an erroneous conception of translation. It does not matter whether a machine or human does the translating—the major considerations are the text to be translated and the communicative situation of the translation (i.e. text type and function of both ST and TT). As a result, there is no one situation that can be used as a point of comparison for HT and MT. Each has its own strengths and weaknesses. Rather than comparing the quality of MT output to HT, it should be assessed on the basis of conventional rules and norms of comparative target language texts—i.e. on a functional basis. Thus, two MT systems can be compared on a functional basis to see to what extent they satisfy end-users' (clients') expectations. As there are different types of MT systems with varying quality and claims to adequacy, there is a need to identify the situations and text types for which these MT systems can perform optimally.

It is generally accepted that MT systems will never be able to produce quality translations of literary works. Humans are creative: computers are not. Computers are good at producing a large volume of output very rapidly. MT systems also have a greater reliability of performance (the same MT system will always translate the same text the same way), but lack intuition (the same MT system will not produce variations in style or lexis unless prompted). Humans cannot compete with computers for speed of output (Arnold 2003: 120). One of the major drawbacks of MT is that it cannot handle new linguistic phenomena (i.e. if the item to be translated is not in its dictionary) like human translators can.

Since computers are very good at following rules precisely and literally, there seems to be some hope for accurate MT as computer coding becomes more robust and better able to handle linguistic transfer. However, MT is not and cannot be creative so it is doubtful that it will ever be capable of dealing with style. This does not mean that MT is not capable of being idiomatic: style is not the same thing as being idiomatic. Idiomatic or fixed expressions are only one type of idiomaticity. There are less "fixed" elements (i.e. collocations) that still come into play if a text is to be considered idiomatic. Both fixed expressions and collocations can be coded into transfer rules of MT systems.

The types of texts that MT can better handle are those that have a rather narrow or restricted style. According to Loffler-Laurian (1996: 46) the only types of text that can be effectively translated by computers are specialized and technical texts. Quite often these texts are required to be translated by law. This is the case in Canada: texts produced in the federal government must be available in both official languages as set out in Canada's *Official Languages Act*. Typically these texts are read for information purposes, rather

than for pleasure, so style is often a secondary factor in such texts. Also, these kinds of texts always seem to follow the "rules" (though the rules may change from text type to text type), whereas in creative texts, sometimes the rules are deliberately broken in order to achieve some kind of literary effect.

1.2.2 Cost of carrying out evaluation

Interestingly, the cost of carrying out evaluations does not seem to be raised in the literature on HT evaluation that I consulted. Perhaps this is because most of this literature seems to be theoretical in nature and tends to deal with literary translation where profitability does not seem to be a major concern. Furthermore, the issue of comparative evaluation (and its cost) does not seem to exist. Perhaps this is because two or more translations of the same literary work are not generally commissioned at the same time and multiple translations of any other type of text are almost *never* commissioned.

However, in the field of commercial, non-literary translation—the area where MT can be and is used—the cost of carrying out evaluations is an important factor. It appears that evaluations are generally carried out to compare MT to HT, the idea being that HT is of high quality whereas raw MT output is not. While almost all human translations are revised (see section 1.1.1.2), the quality of human translations is not generally evaluated in the workplace, the exception being for novice and student translators.

Using a computer as a translation tool impacts the translation process fundamentally. A translation process incorporating an MT system has many aspects that can (and should) be evaluated, which makes cost an important factor of evaluation. Due to the fact that there are many aspects of an MT system that can be evaluated, including

evaluating a system still in development, there are more types of evaluation that can be carried out, depending on the different stakeholders of MT (see section 0.1). These types of evaluation include feasibility tests, internal evaluations, declarative evaluations, usability evaluations and operational evaluations.[13] The tools, tests and procedures used for these types of evaluations can be very time-consuming and require a lot of resources to apply, and can therefore be very expensive.

1.2.3 Issue of what should be evaluated and how

According to Sager (1994: 264-265), MT evaluation can also be carried out at two levels and for two main purposes, as is the case for HT evaluation. These two levels are macro- and micro-evaluation. Macro-evaluation of MT is used to assess the relative merits of one MT system to another (in the case of HT, it is used to compare translations produced by different people or methods, which is almost never done as discussed in 1.2.2), whereas micro-evaluation is used to asses the performance of individual systems (in HT, micro-evaluation compares different translation options or evaluates a single translation). Macro-evaluation is used to judge the extent to which the MT system meets the requirements of its users and is therefore an appreciation of its performance and not an analysis of its potential improvement (Van Slype 1979: 56). Micro-evaluation can be used for several purposes. For example, it can be diagnostic (i.e. look at the reasons for errors) or prognostic (i.e. assess errors for potential solutions and future avoidance). Micro-evaluation is often used during development of an MT system.

[13] For a more in-depth discussion on various types of evaluation and their purposes as well as the various stakeholders in MT, see White (2003) and White (2000b).

Although there are many aspects of MT and therefore many types of evaluation that measure different aspects of MT as discussed above and in section 1.2.2, this thesis will focus on the issue of evaluation from the standpoint of the user-translator and will therefore focus on macro-evaluation. Yet, even from the standpoint of the user-translator, there are many aspects that can (and should) be evaluated ranging from the cost effectiveness of the system, its user-friendliness to the quality of the raw output, etc.[14]

Loffler-Laurian (1996: 68-69) states that many different evaluation measures for MT quality of varying types (e.g. poor quality, intermediate quality) have been developed. Many of these measures are objective in that they attempt to quantify the results. They include evaluations based on the number of errors in the raw output and on the type and amount of post-editing needed to produce a client-ready translation from the raw output. Furthermore a methodology to evaluate style in MT has been developed (Loffler-Laurian 1996: 84-86).

Many of these of these MT evaluation systems use statistics to indicate the quality of the output, including assigning a percentage rating to the translations or assigning a score based on a pre-defined scale. For example, Fernández Guerra (2000: 99-116) has evaluated the output from three MT systems on the basis of linguistic performance (i.e. quality of output), rating the output as percentages of errors and making general statements about the overall quality. Fernández Guerra concludes by stating that MT output is not a finished product (i.e. it is not client-ready) and requires post-editing and polishing (2000: 115-16). However she gives no definition or explanation for polishing,

[14] For a more detailed discussion of other criteria and how they may be evaluated, see Arnold et al. (1994), Fernández Guerra (2000: 95-98), Trujillo (1999) and Van Slype (1979).

which begs the question of whether or not polishing can be considered to be part of the post-editing process.

Loffler-Laurian (1996: 68-69) points out that one of the problems with using percentages is that sometimes it is hard to say exactly what they mean. If for example, a translation is rated as being 80% correct, does that mean that 80% of the MT sentences look like human sentences? Or does that mean 80% of the words are translated with a correct equivalent? Or has translation been judged acceptable by 80% of the people consulted? In any evaluation method, there are always criteria that are ultimately subjective. However, the best judge of any translation is the end-user. For the purposes of this thesis, the end-user also includes the translator who produces the translation based on the client's specifications and therefore is in a position to judge whether or not a translation is acceptable.

Many different scoring scales have been designed to apply to various aspects of output being scored. For example, ALPAC applied a 9-point scale for intelligibility[15] whereas Arnold et al. use a 4-point scale (Arnold et al. 1994: 170). Trujillo uses a 7-point scale to evaluate accuracy (1999: 259-260).

Perhaps the application of scoring scales and percentages is a result of the influence of computer scientists' research on MT; that is, MT started out as a computer science/technology field with little or no linguistic input. This could also be linked to the ultimate goal of automating the evaluation process (White 2000b), which would require a more quantifiable approach since a computer would be doing the evaluation and

[15] ALPAC is the Automatic Language Processing Advisory Panel that was formed by the United States government in 1964 to evaluate the state of MT research. It should be noted that although ALPAC actually applied this 9-point scale to human translations, it was designed in response to the need to evaluate MT output (Arnold et al. 1994: 170). It should also be noted that other authors, including Van Slype (1979: 63-4), indicate that the ALPAC scale is actually a 10-point scale with values ranging from 0 to 9.

computers can really only process mathematical scores. Automatic evaluation could be applied to Example-based Machine Translation (EBMT), a relatively new approach to MT developed over the last 10 years. In EBMT approaches, the MT system breaks down, generalizes and stores previously translated texts to train the system, using these stored items for new translations. One of the features of these systems is their matching function, which lends itself to automated evaluation[16] wherein the MT-generated structures would be compared with those of a proposed target model.

1.2.4 Criteria for evaluating quality of raw output

While many criteria have been identified for macro-evaluation as discussed in section 1.2.3, not all are relevant for evaluating the quality of raw output. Based on my review of the literature, most researchers focus on the two criteria of accuracy and intelligibility (e.g. Arnold et al. 1994 and Trujillo 1999). Hutchins and Somers (1992: 163-64) also use these two categories in addition to style, stating that assessments of style are subjective but appropriateness of style is an important factor. However, as stated in section 1.2.1, style is not typically considered to be particularly important for informative translations, assuming that they are intelligible (which includes the requirement that they be idiomatic). All of these researchers add that error analysis is an important tool for indicating the improvability of a given MT system; however, improvability is not a relevant factor for evaluating the quality of the raw output.

Furthermore, some researchers (e.g. Arnold et al. 1994: 169-70) propose that only intelligibility need be rated for certain purposes since there seems to be a correlation

[16] For a more detailed discussion of EBMT and its matching functions, see Somers (1999).

between accuracy and intelligibility: their research indicates that if a text is intelligible, it is likely to be accurate. I feel that translators will not accept this position due to the fact that one of the major canons of translation is that the meaning must be transferred and in order to do this, the ST must be compared to the TT for accuracy.

As discussed in 1.1.1.4, although quantitative evaluation methods have the strength of being more objective, they do not look at the overall or macro-structure of the text. Indeed, according to Sager (1994: 260-61), due to the fact that MT systems (specifically the artificial languages they generate) are usually conceived at the level of sentences, they are inefficient in making distinctions at the level of larger structures (i.e. paragraphs) or text types. It is my contention that if a metric for coherence were developed, it would have the advantage of being a step closer to measuring the macro-structure of the text, while still being relatively objective.

However, based on the results of my review of the literature, very little research seems to have been conducted in the area of evaluating coherence in MT. In his seminal report on MT quality evaluation, Van Slype (1979: 78-79) indicates that although one researcher (Y. Wilks) considered coherence to be a metric required for measuring quality, he did not indicate how it is possible to rate the coherence of a text. Wilks indicated that the coherence of the TT will be related to and limited by the coherence of the ST. Melby and Warner (1995: 22-23) seem to back up this point of view when they state that MT systems such as Systran recognize and preserve elements that indicate the logical structure of a text such as paragraphs, sections and chapters, thereby preserving the logical structure—and to a certain extent, the coherence—of the ST. Melby and Warner (1995: 62) also state that another way the coherence of STs may be preserved in TTs

when dealing with language for special purposes (LSP) texts is by following the formal structure and order of units in the ST. They argue that technical translations could be specified to correspond at the sentence level, which would tend to retain the ST coherence in the TT. However, they add that preserving the ST structure at the sentence level works best for closely related languages such as English and French (Melby and Warner 1995: 166). In the case of language pairs that are not closely related such as Japanese and English, massive structural adjustments are required.

Krings (2001: 241-42) also discusses the issue of how to create coherence within a text and the need to evaluate it, although he is considering MT evaluation as part of the post-editing process. He feels that factors that create or add to coherence include use of anaphora (pronoun reference), synonyms and paratextual devices which include tables and graphs, etc. Krings advocates the use of what he calls *propositional coherence formation*: when identical terms in the source text refer to same concept, this creates coherence through repetition. This process also applies to propositions. Other structures are counted as coherence formation processes if the text recipient recognizes the repetitive use of a particular structural concept at different places in the text, such as parallel syntactic structures within a list of directions. However, he does not develop any metrics to measure coherence.

Other translation researchers dealing with coherence in translation (but not in MT specifically) include Gerzymisch-Arbogast and Papegaaij and Schubert. Gerzymisch-Arbogast (2001) discusses using the parameters of coherence, thematic and isotopic patterns as potential criteria for measuring translation quality. However, she gives no concrete (quantifiable) metrics for measuring these factors, or even any examples.

Papegaaij and Schubert (1988: 94-98) also discuss coherence and propose analyzing thematic progression to discover coherence patterns in texts. Again, no metrics are proposed to measure the quality of coherence.

Thus it appears that much research remains to be done in the area of developing a metric for coherence. In summary, three important criteria that I feel need to be measured for the evaluation of MT output quality are accuracy, intelligibility and coherence, and the metrics that I propose will be discussed in section 2.4.1.

1.2.5 Post-editing

Given the current state of the quality of MT system output, it typically needs to be post-edited unless it is intended for "gisting" purposes only. Indeed, revision (or in the case of MT, post-editing) is a normal part of the translation process regardless of whether it is done by a human or a computer. However, in the case of post-editing MT output, there are additional issues involved. They include determining the skill set post-editors need and their attitude towards MT and post-editing, the extent to which MT output needs to be post-edited, and the effort or cost of post-editing. Post-editing effort may also be used as an indirect method of measuring the relative quality of MT output.

1.2.5.1 Post-editing is a normal part of the process

In the foreword to Krings (2001: 1), Koby states that although much research has been done in the field of MT and designing MT systems that can produce fully automatic high-quality translation (FAHQT) over the last 50 years, relatively little attention has been paid to the post-editing process. However, this situation is starting to change,

perhaps because it has become evident that the goal of FAHQT is not readily achievable. Generally speaking, post-editing is required but the decision as to whether or not post-editing is carried out and to what extent depends on the quality demanded by the end user. Furthermore, the extent of post-editing required also depends on the quality of the raw output (Austermühl 2001: 165-66).

Koby (cited in Krings 2001: 3) adds that since human translations are normally proofread and edited (or revised) by a second and sometimes a third person, it is unrealistic to expect a computer to produce a translation that does not need to be proofread and edited. However, this expectation of FAHQT requiring no editing was widely accepted in the 1950s and 60s and was even included in the ALPAC report.

As stated in 1.1.1.3, post-editing is the most recent form of revision (Horguelin and Brunette 1998: 5) and as such can be considered as a normal part of the translation process. Horguelin and Brunette add that post-editing is also the most basic form ("la forme la plus brute") because it is often the first human intervention in the MT process. As a result, post-editing is not the same as traditional revision even though there are similarities. The criteria of what needs to be revised or changed can be different. As mentioned in section 1.2.1, texts translated by MT systems are often technical or specialized in nature, having the primary function of transfer of information (i.e. they are informative), so that repetitive language is not considered to be a defect (i.e. style is not of the utmost importance). Furthermore, some MT output will be for internal use only with the criteria that the text be accurate and understandable with the result that style will be even less of a factor. Thus, the extent to which texts need to be edited varies. Loffler-Laurian (1996) introduced the concept of rapid or partial post-editing versus conventional

or maximal post-editing. Maximal or conventional post-editing requires "total" revision, with the result that the TT resembles an original TL text. On the other hand, the objective of rapid post-editing is to change only what is necessary in order to respect the syntactic norms or standards of the TL and to ensure that the meaning is accurate compared with the ST (Loffler-Laurian 1996: 94).

1.2.5.2 Skill set of post-editor

As noted in section 1.2.5.1, post-editing is not the same thing as revision. As discussed in section 1.2.1, Sager contends that computers do not produce human language, but rather artificial language that post-editors must convert to natural language. Koby (cited in Krings 2001: 4-7) agrees and adds that human translators construct meaning from the sentences they read in the ST, then express this meaning in the TT. These translators also have knowledge of the SL and the TL as well as both cultures. Human revisers also have this knowledge. Computers process text differently: computers do not "know" both languages nor do they have any cultural knowledge. However, MT systems can still be used to produce draft translations, which are then submitted to post-editors who revise them into a client-ready state. Accordingly, the task of the translator (in this case the post-editor) is changed. Since the task is not the same, the issue arises of what exactly comprises the skill set of post-editors and how it compares with the skill set of traditional revisers.

Koby (cited in Krings 2001: 1-2) states that Krings investigates the argument that due to the fact that a computer is doing the translating, post-editors do not need to be translators or even be bilingual. As part of his investigation Krings examines the issue of

whether or not individuals can post-edit the raw output texts without referring to the ST. Koby adds that this position turns out not to be a viable alternative. It is difficult to see how accuracy could be ensured if there is no reference to the ST. Krings found that post-editing involves three texts: the ST, the MT output and the TT (post-edited text) (2001: 166-167). Accordingly, post-editors must have translation skills, or be bilingual at the very least.

According to Schäfer (2003: 187), generally the post-editor and the translator are the same person, which can to lead to situations where the post-editor/translator expects the same quality output from the MT system as from HT. This type of situation only adds to the widespread misperceptions of MT systems and their ability to produce FAHQT. Schäfer adds that according to Arnold et al. (1994: 6), MT systems have not been designed to translate Shakespeare (i.e. they will not produce elegant prose). As a result, it is important for post-editors to have a positive and open-minded attitude. Accordingly, post-editing demands a certain degree of tolerance and the ability to draw clear boundaries between purely stylistic improvements and linguistic corrections (Koby cited in Krings 2001: 16).

Schäfer (2003: 187) states that MT texts are linguistically different from HT texts. Consequently, post-editors require experience and skills in recognizing typical machine errors. MT output errors tend to be typological and recurring (Koby cited in Krings 2001: 7). In other words MT output typically contains more errors, but they are repetitive. These errors can be easier and more irritating to fix due to the fact that they tend to be errors that humans would never make. While human translators may give the wrong slant to a whole sentence or even longer passages, machines make more local mistakes, often

confined to a single lexical item or syntactic structure. However, many errors or mistakes in MT output are not immediately apparent and the text could appear comprehensible at first glance (Schäfer 2003: 187). Accordingly, post-editors must compare the MT output to the ST to identify errors resulting from incorrectly analyzed syntactic structures or from defects in the ST.

In summary, in addition to the obvious skills of keyboard and computer knowledge (Koby cited in Krings 2001: 16), post-editors must be bilingual (they are usually translators), have a positive and open-minded attitude and have the ability to recognize typical MT errors.

1.2.5.3 Issue of how much post-editing needs to be done

The research in this area seems to deal mainly with the issue of the "too much" versus "too little" problem, which stems from the fact that post-editing is what Schäfer calls the "unknown task" (2003: 186). He adds that as a result of the MT projects at SAP (a large software company), a common or universal post-editing guide for all translators is being developed. Underlying this guide is a systematic error typology that serves as a methodological framework for the research on post-editing. The purpose of this guide is to give translators instructions about the "unknown task," to familiarize them with its linguistic implications resulting in a further open-minded attitude towards MT, and to increase translators' efficiency since familiarity with error patterns will reduce post-editing time (Arnold et al. 1994: 173-175).

Loffler-Laurian (1996: 82-84) makes a distinction between correction and post-editing stating that if raw MT output is given to translators and represented as human

translations, then the translators will "correct" the output. If translators are aware that the texts are raw MT output, then they will post-edit them. As a result, they will tend to experience less frustration at unintelligent and repeated errors. Post-editing is a technique and should not be viewed as a creative or literary task.

The scope of post-editing required depends on the quality of the raw MT output, the text type and the purpose of target text (Schäfer 2003: 187). Other factors impacting the scope of post-editing include the end-user or client, volume of documents to be translated and the timeframe/deadline (Allen 2003: 301). Loffler-Laurian feels that end-users are the best judges of the degree of translation quality required because they know what the texts will be used for (1996: 69).

Darbelnet has developed a set of seven criteria for revision that can be used as the basis for a methodical approach to revision (Horguelin and Brunette 1998: 34-35), and which can also be applied to post-editing. Darbelnet states that of the seven levels of criteria (semantic, idiomatic, style, cultural, allusional (e.g. film titles), "interior" level (i.e. when an implicit element needs to be explicated)), only the first two levels may need to be taken into account for technical or specialized translation. Thus according to Darbelnet, the two main criteria for revision and by extension post-editing are semantic accuracy (meaning) and idiomaticity (intelligibility). When dealing with technical and specialized, informative texts—as is the case with MT—revision or post-editing for other criteria such as style is not always essential.

Thus the post-editing focus is on adjusting MT output so that it accurately reflects the meaning of the original text and is idiomatic. This approach should not pose a problem for specialized texts produced in a government context as their purpose is to

inform: they are not literary documents whose style is an important factor. According to Hutchins (quoted by Koby in Krings 2001: 9) the standard of quality that human translators apply to human-translated texts is that everything must be perfect. However, end-users do not seem to be so exacting—there is now a growing realization that that for many end-users, style refinements are not necessary so long as the text is accurate and comprehensible. This is difficult for many translators to deal with and presumably at least partially explains a lot of their resistance to MT.

1.2.5.4 Criteria for evaluating post-editing effort

Because computers can translate huge volumes of text quickly and cheaply, one of the major costs for producing client-ready MT is post-editing. Since post-editing represents such a large portion of the cost of producing usable MT output, it makes sense to use post-editing cost or effort as the metric (albeit an indirect one) to evaluate the relative quality of MT systems' output. However, this brings up the issue of what this metric should consist of and how it should be measured (Krings 2001: 532).

Krings differentiates three types of post-editing effort: temporal, cognitive and technical. Temporal post-editing effort can be defined as the amount of time taken to post-edit a translation: it is also the most externally visible and economically significant aspect of post-editing effort. Technical post-editing effort can be determined by calculating the number of physical changes made to the raw output using either key strokes or number of insertions and deletions. Cognitive effort can be defined as the mental effort required by a post-editor to edit the MT output.

According to Krings (2001: 55), technical and temporal efforts are externally observable aspects of post-editing and can therefore be measured. Cognitive effort is not externally observable and is therefore more difficult to quantify. For the purposes of this thesis, externally observable and quantifiable (and therefore comparable) aspects of post-editing are relevant. Cognitive effort *per se* is not relevant, although it is indirectly considered due to the fact that cognitive effort takes time and can be assumed to be included in temporal effort. Furthermore, although technical effort is potentially a viable measure, it is not as easily measured as temporal post-editing effort. Accordingly, temporal post-editing effort (i.e. total time taken) seems to be an obvious choice as the metric for measuring post-editing effort and as an indirect measure for the quality of raw MT output.

According to Krings (2001: 553) the evidence suggests that more post-editing processes (i.e. amount of post-editing required) are necessary not for poor machine translation, but rather for medium-quality translations (i.e. there is a non-linear relationship between MT output quality and post-editing effort). Krings' reasoning is that medium quality MT output requires extensive comparison between the MT output and the ST to determine which elements of the MT are acceptable and which are not. However, his conclusions have not been tested extensively. Furthermore, the sole measure of MT quality is a sentence-by-sentence rating (5-point scale) done by post-editors who are not given specific criteria; they simply rate quality according to their own personal judgment (Krings 2001: 251). Another potential shortcoming in Krings' findings stems from the fact that he seems to assume that unless MT quality output is obviously

poor, the MT output is not compared to the ST, which has ramifications for assuming the output is accurate.

However, not all MT researchers agree with Krings' findings. For example Fernández Guerra (2000: 35) states that the amount of post-editing depends on the quality of the output. It should be noted that the term quality is very subjective, since the amount of post-editing required also depends on the purpose or function of the translation and the time available to prepare it (i.e. the deadline or timeframe).

For the purposes of this thesis, time will be the metric for measuring post-editing effort for each of the MT systems being compared. However, post-editors need to be aware of the distinction between changing only what is necessary to ensure accuracy, intelligibility and coherence, and imposing their own preferences and style on MT output, thereby unnecessarily taking more time and increasing the cost. Revisers generally do not impose their style on human translators, so it seems logical that post-editors should not do the same to MT output. A more detailed discussion of the post-editing metric used in this thesis is contained in section 2.4.2.

1.3 Concluding remarks

As van der Meer sums up the state of translation automation (2003: 183-84), 50 years of research and development in the field have generated very few results. The key inhibiting factor has been the lack of a quality standard. People are emotional about language and instinctively want maximum quality—but are unable to measure it. However, he feels that this situation is about to change; the huge increase in demand for translation and the drive towards globalization will lead customers to adopt simple and

standardized quality metrics to measure the quality of translations. The purpose of this thesis is to work towards developing simple, effective quality metrics (i.e. accuracy, intelligibility and coherence as direct metrics, and post-editing effort as an indirect metric) and by extension a straightforward quality evaluation methodology that can be applied to MT systems.

Chapter 2 Preliminary Design of the Evaluation Methodology

This chapter will describe the preliminary design of the evaluation methodology, which will then be applied in a pilot study described in Chapter 3.

The EAGLES[17] Evaluation Working Group (EAGLES 1999) has developed a seven-step recipe or methodology to carry out an evaluation of any language technology system, including MT systems. The EAGLES Evaluation Working Group methodology consists of the following seven steps:

1. State the objective or why the evaluation is being done.
2. Elaborate the task model.
3. Define the top-level quality characteristics to be evaluated.
4. Produce requirements for the system being evaluated.
5. Devise the metrics to be applied for the requirements.
6. Implement the evaluation: design the execution of the evaluation.
7. Execute the evaluation and draw conclusions.

I have chosen to use this recipe as a guideline for developing an evaluation methodology for MT systems, and I will describe how I applied these steps to my own research.[18] Note that in some cases, there is a certain degree of overlap between information presented in the various sections. While I have no desire to be redundant,

[17] EAGLES is the Expert Advisory Group on Language Engineering Standards (http://www.ilc.cnr.it/EAGLES96/home.html).

[18] Steps 1 to 5, which apply to the design of an evaluation methodology, will be described here in Chapter 2, while steps 6 and 7, which pertain to the implementation of the methodology, will be described in Chapter 3.

sometimes it is necessary to repeat certain points in different sections for the sake of clarity.

2.1 Objective

The objective is to design and test an evaluation methodology that can be used to compare the parallel output of various MT systems and state which is better.[19] In addition, total post-editing effort will be calculated based on time and used as a metric to measure the relative quality of the raw output of both systems being tested.

2.2 Task model

For the purposes of the evaluation methodology being developed and tested, I will choose two relatively inexpensive MT software systems that are readily available to freelance translators and use them to translate the same source text. I will compare and score the quality of the raw output of the two systems (i.e. a comparative, declarative evaluation).

I will then post-edit the raw output of both systems, keeping track of the amount of time taken to post-edit the texts to the point where the quality is good enough to deliver to the client (i.e. client-ready). It is important to note that these texts will be comparable only in the sense they are high quality translations that would be acceptable to the client. The fact that translations of the same source text are comparable does not

[19] Speed and cost of producing the raw output will not be considered as factors for the purpose of this thesis because they will be roughly the same for both systems: as explained in section 0.3.2.3, these were criteria that were used when selecting the MT systems to be tested.

necessarily mean that their wording is similar. The premise is that the better quality text will take less time to post-edit—again, a comparative, declarative evaluation.

The evaluation methodology will be refined based on the results of the preliminary testing (Chapter 3). The refined methodology will be tested (Chapter 4) using a wider variety of specialized texts that will be translated by the same two MT systems used for the preliminary study. Each set of raw output texts will be scored by an independent evaluator who will be using the refined methodology. The sets of raw output for each ST will also be post-edited by the evaluator.

The results will be examined and conclusions drawn as to whether or not the evaluators were able to declare which system produced the "better" output for the task in question. Additional refinements or suggestions for future research may be made.

2.3 Definition of top-level characteristics to be evaluated

For the evaluation of the quality of the raw output of the two MT systems, accuracy, intelligibility and coherence have been defined as top-level characteristics to be measured. Post-editing effort has also been defined as a top-level characteristic to be measured.

Generally speaking, most translation quality evaluation systems for MT measure accuracy and intelligibility, two characteristics or criteria that I agree are essential for high quality translations. Although no translation quality evaluation methods for MT appear to measure coherence (see section 1.2.4), I feel that it is also a characteristic of a high quality translation just as it is a characteristic of any well-written text. In his seminal work on MT evaluation, Van Slype (1979) has suggested that coherence should be

evaluated in addition to intelligibility and fidelity (accuracy).[20] Accordingly I have decided to include coherence as a characteristic to be evaluated.

An accurate text, sentence or translation unit is one whose meaning is correctly conveyed in the target text (White 2003: 216). An accurate translation, regardless of its size, will have no meaning errors or omissions.

I have chosen to use the sentence as the unit of evaluation for accuracy. Most evaluation systems for MT also use the sentence as the unit of evaluation (e.g. Trujillo 1999, Nagao et al. (cited in Trujillo 1999), Carroll (cited in Van Slype 1979), and Van Slype 1979). Although many researchers feel that translation units (TUs) should be used as the basis for evaluation of accuracy, there is some debate as to what exactly constitutes a TU as discussed in section 1.1.2. Some researchers feel that TUs consist of propositions (Krings 2001), while others feel they are smaller (Horguelin and Brunette 1998). Furthermore, the segmenting of texts to be evaluated into TUs is a cumbersome and time-consuming process which may not be justifiable for the results obtained. Since MT systems translate on a sentence basis and, therefore, each ST sentence is rendered by one TL sentence regardless of the MT system used, the relative scores will be comparable as each TL sentence should contain the same number of TUs.

Intelligibility, or "target-language fluency" (White 2003: 216) can be defined as the ease of comprehension by the reader of the text. To be of any use, a text must be intelligible. As is the case with accuracy, intelligibility is generally rated at the sentence level, again most likely for ease of scoring (e.g. Trujillo 1999, Nagao et al. (cited in Trujillo 1999), and ALPAC 1966).

---

[20] Unfortunately, Van Slype did not provide practical guidelines as to how coherence could be measured. The approach that I have designed to measure coherence will be described in detail in section 2.4.1.3.

Coherence is also a measure of target-language fluency or intelligibility, but it focuses on a larger unit. A coherent text is one that hangs together and whose organization makes sense and can easily be understood by the reader. By definition, coherence is a measure of a series of sentences (paragraph) or an entire text as opposed to a single sentence or translation unit. Based on a thorough review of the literature, I found that to the best of my knowledge, there is no precedent for coherence testing/scoring in MT. As mentioned previously, Van Slype (1979) included coherence as a criterion for quality assessment as certain researchers had considered it in their proposed methodologies. However none of these researchers actually devised a method for measuring coherence, perhaps because MT uses the sentence as the base unit for translating, whereas coherence is a characteristic of larger units (i.e. a paragraph or even an entire text). Therefore, the inclusion of coherence testing in this evaluation technology represents an original contribution. I feel it is important to test for coherence because translation is about creating a text (as is all writing), not merely a string of sentences (see, for example, Fromkin et al. 1997: 144-151).

In fact, one of the criticisms often leveled at both MT and CAT systems is that translations are done at the sentence level with no thought given to the text as a whole (Macklovitch and Russell 2000: 137). One popular type of CAT tool is known as a translation memory, which "recycles" sentences and sentence fragments. The result can be a "sentence salad" that does not necessarily hang together as a coherent text (Bédard 2000). Devices such as omissions, ellipses, pronouns and other anaphoric and cataphoric devices, etc. help streamline a text and make it easier to read. Although these devices likely exist in a source text (i.e. it is coherent), the means of conveying them in the TL do

not necessarily coincide with the means used in the ST. In other words, different cohesive devices may be used, and these may appear in other sections of the target text. I have therefore chosen to evaluate coherence at both the paragraph and text level.

Post-editing time has also been defined as a top-level characteristic, the idea being that the less time it takes to post-edit a translation, the better the quality of the raw output. Since MT output is not typically good enough to be delivered to clients without revision, post-editing plays a key role in the MT process, as described in section 1.2.5. The major cost associated with post-editing is the time required to edit the target text. The hypothesis that will be investigated in this thesis is that there is likely a correlation between post-editing times and raw output scores: the higher the quality of the raw output (based on scores given for accuracy, intelligibility and coherence), the less time will be needed for post-editing.

Post-editing is a tricky and sensitive issue. The extent to which a text needs to be post-edited and the decision as to what exactly qualifies as an error will vary from translator to translator (or post-editor), and perhaps also from client to client. Documents for internal use will not likely need to be as extensively post-edited as outbound (i.e. for publication) documents. Allen (2003: 301-307) differentiates between three levels of post-editing: rapid (i.e. for internal documents), and minimal, and full post-editing (i.e. for external documents). However, although fully post-edited texts will be idiomatic, they are not edited for style. The difference between what is a necessary revision (i.e. to make a text idiomatic) and what is stylistic is very subjective. For the purposes of this thesis, post-editors will be instructed to edit so that the texts are grammatically correct and idiomatic and to make no changes for style or personal preferences. The job of the post-

editor is limited to fixing problems within existing structures, and s/he will not be given the freedom to completely rewrite passages as this is no longer post-editing, but rather imposing personal style and preferences on the "translator" (in this case the MT system). The post-editor is not translating the text and as a result it will not reflect his/her style.

2.4 Metrics to be applied to output of MT systems

The metrics that will be applied to the raw MT output are based directly on the top-level criteria described in section 2.3. In other words, the relative quality of the raw output will be measured both directly (i.e. by scoring for accuracy, intelligibility and coherence), and indirectly (i.e. by measuring time required for post-editing).

2.4.1 Metrics to be applied to raw output

The three criteria used to score raw output are accuracy, intelligibility and coherence. Any variations (i.e., subjectivity) in the way various individuals (translator-users) score raw output will not adversely affect the comparative scores because the same individual will score both texts (raw output), so that his or her personal preferences will be reflected in both scores. Moreover, since the idea is that the individual/translator doing the scoring is the person who will be purchasing and using the MT system, it is reasonable that the scores reflect his or her preferences.

2.4.1.1 Accuracy

For the purposes of accuracy, a 4-point scale is used. This scale is loosely based on a scoring grid used by some revisers at the Government of Canada's Translation Bureau to evaluate student translations. This scale is divided into major and minor meaning errors. Accuracy is scored at the sentence level and compares the translation (raw output) with the original document (source text). For the purposes of the preliminary testing, accuracy is evaluated at sentence level as this is the easiest and most-straightforward unit to use as discussed in section 2.3.

Sentences will be assigned a score ranging from 1 (no meaning errors) to 4 (major meaning errors). Accordingly, the lower the score a text receives, the more accurate it is deemed to be. The 4-point scale for accuracy is as follows:

1. completely accurate: no meaning errors.
2. accurate: minor meaning errors (no more than 2), especially words/terms not in dictionary.
3. adequately accurate: minor meaning errors (2 to 5) including minor omissions and verb tense errors. Errors do not fundamentally change meaning.
4. inaccurate: major meaning error including major omission or more than 5 minor meaning errors. Error changes fundamental meaning.

Table 1 provides some examples of minor and major meaning errors.

Error type	Example	Explanation
Minor meaning error:	ST: fournir des **conseils** au fonctionnaires fédéraux MT: providing **councils** to the federal civils servant	The noun **conseils** has been translated as **councils** instead of **advice**. The meaning can still be understood.
Major meaning error:	ST: en ce qui **a trait** au SRAS MT: in what **milked** with the SRAS	The verb **traiter** (to deal with) has been incorrectly parsed as **traire** (to milk). The meaning has been fundamentally changed.

Table 1 – Examples of different types of meaning errors

2.4.1.2 Intelligibility

The scoring scale for intelligibility has been adapted from the 5-point scale designed by Trujillo (1999: 258). For this scale, intelligibility is scored at the sentence level and is scored referring only to the raw MT output (i.e. the output is not compared with the source text).

Sentences will be assigned a score ranging from 1 (no editing required) to 5 (totally unintelligible; sentence must be completely rewritten). Accordingly, the lower the score a text receives, the more intelligible it is deemed to be. Note that sentence length and therefore number of errors may vary. However, the seriousness of the errors can be seen to compensate for this. Accordingly, the minimum/maximum numbers given for intelligibility are guidelines rather than hard-and-fast rules.).

The 5-point scale for intelligibility is as follows:

1. completely intelligible: no rewriting or editing required.
2. intelligible: some problems with usage or word order (up to 3 errors).
3. adequately intelligible: basic meaning can be understood but some details are not clear and sentence contains a maximum of 5 errors of usage or word order or other grammatical errors.
4. unintelligible: reader can only guess at meaning; many and more serious grammatical errors than for category 3 (i.e. more than 5 errors of usage and word order, but not more than 10).
5. completely unintelligible: more than 10 errors of usage and word order or serious grammatical errors that make the sentence unintelligible.

Table 2 provides examples to illustrate the differences between the degrees of intelligibility.

Intelligibility category:	**Example**	**Explanation**
1. completely intelligible	Consequently, Health Canada considers that the custom{*usage*} of masks is not necessary.	There are no usage or word order errors in this sentence. **Custom/usage** is a meaning error and is taken into account in the accuracy scoring.
2. intelligible	**The** employers can use this information and **these** advice in order to determine measurements of health and safety which they consider suitable for their employees.	The articles **the** and **these** need to be deleted. (2 errors) **Measurements** is a meaning error and taken into account in the accuracy scoring.
3. adequately intelligible	Q. Why Health Canada **does not he** recommend the custom{*usage*} of masks by **the** federal civil servants such as **the** customs officers in **airports Pearson** and **of Vancouver**?	Word order errors include **does not he** and **airports Pearson** and **Vancouver**. The articles **the** (federal…) and **the** (customs…) as well as the preposition **of** need to be deleted. (5 errors) **Custom/usage** is a meaning error and taken into account in the accuracy scoring.
4. unintelligible	Health Canada believes that **the** employees of the **government federal do not risk to become patients**, because **the contacts of**	The articles **the** (employees), **the** (passengers) and **the** (customers)

	the employees with **the** passengers or **the** customers are of relatively short duration and are not considered as narrow.	need to be deleted. Word order errors include **government federal** and **the contacts of the employees**. Verb form errors include **do not risk** and **to become**. (7 errors) **As narrow** is a meaning error and taken into account in the accuracy scoring.
5. completely unintelligible	How Santé Canada **does it** support **the** federal **civils** servant **in what** milked **with the** SRAS?	Word order errors include **does it**. The articles **the** (federal) and **the** (SRAS) need to be deleted. The **s** should be deleted in **civils** and should be added to **servant**. **In what** is a grammatical error that needs to be reformulated. **Santé Canada**, **SRAS** and **milked** are meaning errors and taken into account in the accuracy scoring.

Table 2 – Examples of different types of intelligibility-related errors

The scores for intelligibility are obtained by having the rater/translator read each sentence and assign it a score according to the scale. No other testing methods such as

cloze tests or multiple choice questions are used to arrive at a score. The purpose of this evaluation methodology is to allow individual translators to evaluate MT output for quality so that they may be in a position to decide which system is better for their needs. Accordingly, each translator (or rater) is in the best position to decide what score should be attributed to each sentence. The objective is not to make a statement on the absolute quality of a translation but on the relative quality of each version of the output as produced by different MT systems.

2.4.1.3 Coherence

The criteria and scale used to score the raw output for coherence have been inspired by those developed by Trujillo (1999) for intelligibility. I felt this was justified because intelligibility and coherence are related concepts. Coherence can be defined as the general organization that accounts for the connectedness of a text (Fromkin et al. 1997: 471). Whereas intelligibility is applied at the sentence level, coherence is applied at the paragraph and/or text level. Coherence was scored looking only at the raw output of the MT systems tested (i.e., with no reference to the source text). Coherence was scored at paragraph level as well as for the text as a whole. For example, if a text was made up of 5 paragraphs, there would be a total of 6 scores for coherence – one for each paragraph and one for the overall text.

The 5-point scale for coherence is as follows:

1. completely coherent: passage (i.e. paragraph or text) is logical and consistent. No editing required for coherence.

2. coherent: passage is logical but contains some problems, especially with pronoun reference. No minimum or maximum number of errors has been established. Paragraphs and texts vary in length so that the number of errors that result in the passage being incoherent will vary.
3. adequately coherent: overall meaning can be understood but some sentences (i.e. up to 50%) in passage are not logical.
4. incoherent: reader can only guess at overall meaning: many (i.e. more than 50%) of the sentences in passage are not logical.
5. completely incoherent: meaning of paragraph or document cannot be detected.

2.4.2 Metrics for post-editing

As discussed in section 2.3, post-editing plays a key role in transforming MT output to a state where the quality is of a standard that can be delivered to a client (i.e. client-ready). The major cost associated with post-editing is the time taken by post-editors (who are typically paid by the hour) to revise the raw output.

For the pilot test, the raw output of both MT systems will be post-edited by a single person (the researcher), who will keep track of the time to the nearest minute. The post-editor will have already seen the source text (i.e. the post-editor has also scored the raw output for accuracy comparing it to the ST). Each translated text (i.e. the raw output) will be divided into two parts and the post-editor will edit the first part of text A, followed by the first part of text B. The post-editor will then edit the second part of text B, followed by the second part of text A. This procedure will be followed for fairness: normally it would take longer to post-edit the first text (text A) than the second (text B)

because the content is less familiar. By editing half of each text first, this bias will be reduced. Therefore, the circumstances and conditions for the post-editing of each text are equivalent and should not affect the comparative ratio of post-editing time.

2.5 Summary

In summary, the proposed evaluation methodology is designed to enable evaluators to compare the parallel output of various MT systems and state which one they feel is better. Applying the methodology can be broken down into two main steps as follows:

1. the quality of the raw output from each MT system being tested is evaluated by applying the scoring scales as described in section 2.4.1 for each of the three criteria defined as the top-level characteristics to be measured. As stated in section 2.3, accuracy, intelligibility and coherence have been defined as top-level characteristics or criteria to be measured for the evaluation of the quality of the raw output of MT systems. The total (normalized) scores are compared to see which system is better.

2. the raw output is post-edited by the evaluator who keeps track of the time taken. The total time taken to post-edit the raw MT output (i.e. post-editing effort) is defined as the top-level characteristic to be measured (see section 2.4.2). The total times are compared to see which system is better.

The design outlined here will be applied and evaluated in Chapter 3.

Chapter 3 Initial Implementation of the Evaluation Methodology

The evaluation methodology that was designed in Chapter 2 will now be implemented in a pilot study. The results of this pilot study will then be evaluated with a view to refining the evaluation methodology before testing it on a larger scale (Chapter 4).

3.1 MT system selection

For the purposes of the preliminary testing, two MT systems were chosen based on the criteria defined in 0.3.2.3. As stated in 0.3.3, the two MT systems chosen were Systran Classic (1999 version) and ReversoPro 98. Both systems[21] are commercially available at a cost that most translators could afford (i.e. less than $1,000 CD). Both systems are easily installed on a PC and do not have exorbitant system requirements in that they run on a Pentium processor with 256K of RAM. Both handle the French-to-English language pair and are compatible with Word software.

Systran is considered to be a first generation system as it was originally developed in the 1970s.[22] Although it has been extensively updated and patched many times, Systran is considered to be a direct MT system because it relies extensively on dictionaries and does not perform complete sentence parses (Hutchins and Somers 1992: 177-8). Systran is not based on any linguistic theory (Loffler-Laurian 1996: 38).

[21] Or subsequent versions of these systems.

[22] Peter Toma, the creator of Systran, initially began MT research in the late 1950s, and did some early prototype development of Systran in the 1960s, but it was not until 1975 that the first fully functioning version was developed (Hutchins and Somers 1992: 176).

ReversoPro is a transfer system, as are most commercially available systems. Transfer systems analyze each ST sentence (using a grammar and rules specific to the source language) into an abstract representation, before mapping or transferring it to a corresponding abstract representation of the target language. The representations of the sentences are then generated in the target language (Arnold 2003: 116-7). Transfer systems are language-pair specific and unidirectional.

Both systems come with a general (or main, or stem) dictionary. Specialized and technical dictionaries may be added. In addition, both systems provide the option of adding user-defined dictionaries that can be edited and customized according to the clients' needs. In both MT systems, the specialized/technical dictionaries can be chosen as an additional source when texts are translated. For the purposes of this thesis, only the main dictionary of each respective system was used. No terminology was added to the main dictionary, which was not updated in any way.[23]

3.2 Text selection and pre-processing

For the purposes of the pilot study, the two MT systems were used to produce translations of a specialized (i.e. not technical) text taken from the Internet (Government of Canada Web site). Standard French as used in Canada is the source language of the text chosen to be translated and the target language is standard Canadian English. Factors such as formatting (including fonts) are not used as criteria for the purposes of this evaluation methodology.

[23] Although the performance of both systems could certainly be enhanced by building up their respective dictionaries, doing so would not really affect the *comparative* type of review I am conducting. Accordingly, I have not built up the two systems' dictionaries in any way.

The text chosen was a 354-word French-language text from the Santé Canada Web site on the specialized subject of Severe Acute Respiratory Syndrome (SARS) (see Appendix A). A specialized document was chosen not because experience has shown that "sublanguage" translates better than unrestricted language, but rather because specialized documents are the type of texts **most** Canadian translators actually translate.[24] A great many of the documents that need to be translated in Canada are generated by the federal government and related agencies and organizations, and therefore tend to be of a specialized nature.

The English version of this text (see Appendix B), which was produced by a human translator, was also retained from the Web site and used as a basis (along with the source text) to help evaluate the accuracy of the raw output produced by the two MT systems.

The raw output from the two MT systems appears in Appendix C. Note that the source text document was edited to remove typos before being run through the MT systems so that no errors in the output resulted from input errors. The text was not, however, "pre-edited" for anything beyond typographical errors. In other words, the text was not altered to remove any lexical or structural ambiguities that it might contain.[25]

---

[24] Selection of this text as being typical was based on my experience in university translation practicums, as a freelancer and as an employee of the Translation Bureau.

[25] Text can be pre-edited, making changes with a view to eliminating errors in translation. Generally speaking texts are pre-edited using a controlled language which contains a limited vocabulary with each term having one meaning only and no synonyms. Also, the syntax is very simple. For more detailed discussion on controlled language, see Lockwood (2000) and Nyberg et al. (2003).

3.3 Applying the raw output metrics

For the preliminary or pilot testing, the raw output of both systems was scored first for accuracy, then intelligibility and finally for coherence. The accuracy scoring was carried out using both the source text and a human translation for reference. For the coherence and intelligibility scoring, which are measures of readability, the raw MT output was examined with no reference to either the source text or a human translation. After the scales for accuracy, intelligibility and coherence were applied to both sets of MT output, the scores were normalized by adding up the sentence/paragraph scores and dividing by the total number of sentences/paragraphs. Normalization of the scores is an important step that is necessary to facilitate comparison of the two systems. Due to the fact that all three scales assign lower scores to better performance, the system with the LOWER overall score was deemed to have better quality raw output.

Note that in the context of this experiment, accuracy, intelligibility and coherence are considered to be equally important. I feel this is justified because the goal is to produce a translation that is ready to be delivered to a client, and such a client-ready translation must reflect all three of these characteristics.[26]

3.4 Applying the post-editing metrics

The raw output of the two systems was then post-edited to client-ready quality. As discussed in section 2.4, the fact that each translation must be accurate, intelligible and coherent does not imply that their wording should be the same or even similar.

[26] It should be noted that there are MT applications such as "gisting" where the emphasis is on being able to get the main thrust of the text, with little concern for style or readability, but such applications are beyond the scope of this thesis. For more on gisting, refer to Allen (2003).

Accordingly, the raw output was not edited for style or personal preferences of the post-editor. The total time taken for post-editing to the nearest minute was used as the basis for scoring. The system output that received the lowest time score (i.e. required the least amount of time to post-edit) was considered to have better quality raw output.

As a final step, the scores for post-editing and raw output quality were compared for correlations (i.e. to see if the same system was rated as better in both categories).

3.5 Data analysis and evaluation

Evaluations of the actual data (i.e. scores for accuracy, intelligibility and coherence, as well as the post-editing times) produced during the pilot test are presented in this section.

3.5.1 Raw output scores

A detailed indication of the scores received for each sentence/paragraph/text contained in the two target texts is presented in Appendix D, while the total scores for each of the raw output criteria are summarized in Table 3.

Criteria:	**Systran** Score	Normalized score (= score / total # of units)	**ReversoPro** Score	Normalized score (= score / total # of units)
Accuracy	33	2.50	26	2.17
Intelligibility	34	2.83	34	2.83
Coherence	13	2.17	11	1.83
Total	81	7.50	71	6.83

Table 3 – Scoring results for raw output

Based on these scores ReversoPro, with a normalized total score of 6.83 produced an overall higher quality (better) raw output than Systran, which scored 7.5. It is interesting to note that the ReversoPro scores for all three categories were equal to or lower (i.e. better) than those of Systran. In other words ReversoPro produced "higher" quality output no matter the category or criteria being evaluated.

3.5.2 Post-editing results

It took 19 minutes to post-edit the Systran raw output and 17 minutes to post-edit the ReversoPro raw output (see Appendix E). Based on these results, ReversoPro produced higher quality (better) raw output, which is consistent with results obtained using the scoring of raw output metrics (see section 3.5.1). Thus, this appears to confirm the hypothesis that there is a correlation between post-editing times and raw output scores

in that the higher or better the quality of the raw output, the less time needed for post-editing.

3.6 Refining the methodology

After applying the methodology in the pilot test, I found that for the most part the evaluation system works well. However, I felt that some changes or refinements should be made to improve the methodology and make it easier to apply. The following sections include a discussion on the changes that I considered, an analysis of the situation (i.e. the pros and cons for making the change), and my final decision about whether or not the change would be implemented in the refined methodology.

3.6.1 Weighting of criteria

As noted in section 3.3, the three criteria of accuracy, intelligibility and coherence were all assumed to be equally important to the quality of the raw output. However, a potential problem would arise, for example, in the case where the output of two systems receive the same cumulative score but the first (A) has a good accuracy score and a poor coherence score, and the second (B) has a good coherence score and a poor accuracy score. The question arises as to whether or not it can it be said that these systems are really equivalent in terms of quality and as to whether cumulative scores are really a useful or fair measure. Arnold et al. (1994), among others, make a case for intelligibility being the sole criteria for deciding the quality of MT output because there seems to be a correlation between accuracy and intelligibility: if a text is intelligible, then it is accurate.

Although on the surface it would seem that intelligibility is more important for these researchers, they do not actually deem one aspect to be more important that the other; they merely state that intelligibility scores reflect accuracy scores. I found no cases where researchers tested for coherence.

I feel that accuracy should be more important for translating informative texts, since their main purpose or function is to convey the content (i.e. if reader can understand content, then it does not matter if it takes longer to read). However, it may take longer to post-edit unintelligible text than to post-edit inaccurate text. This is why I have included post-editing time as an additional metric for measuring quality of MT systems output.

In the absence of any conclusive evidence that any one of these three metrics is more important that the others, I have not weighted any of them.

3.6.2 Level of detail in scoring system

The next point I considered was changing the level of detail in the scoring system. Scoring scales with more points would allow finer grading and perhaps make it easier to define fine differences in quality. However, upon reflection, I decided that the scales did not need to be more fine-grained. For example, if a 9-point scale were to be applied to accuracy, scorers would need to spend a lot more time making decisions between small differences in quality. The purpose of this methodology is to allow potential purchasers to form an idea of the relative quality of MT systems when compared to one another, not to score the quality of the output to a fine degree. The major difficulty associated with the scoring system was during the development phase, when it was necessary to decide how many points the scales should consist of and what the differences were between the

points of the scales. Once the initial discomfort at making the decision of scoring gradients was overcome, the resulting scoring system was relatively straightforward to apply.

3.6.3 Order of application of raw output criteria

On the whole, the scoring of the three main criteria of accuracy, intelligibility and coherence seems to be effective. However, I feel intelligibility should be scored first (i.e. before accuracy and coherence) because it requires less effort on the part of the reader/scorer. Once a reader understands the content of a text, it is easier for him or her to verify whether or not this content is accurate. It is not practical to check whether facts are accurate before reading a text and understanding it.

I also feel it would be more efficient to score coherence immediately after scoring intelligibility, as these two criteria are related (they both deal with readability of text). Accordingly, I will change the order of the scoring of raw output so that intelligibility is scored first, then coherence and lastly accuracy.

3.6.4 Post-editing time measurement

There were no apparent problems with the post-editing process. However, because the scores obtained during the pilot system were so close (i.e., 19 minutes and 17 minutes), I did consider the option of measuring post-editing time to the second rather than the minute, as this would more accurately reflect the total post-editing time. It is possible a post-editor could spend the same number of minutes editing both versions of a

text although if measured to the second, the scores would be different. For example two versions that took 20 minutes each to post-edit could actually have taken 19 minutes and 31 seconds and 20 minutes and 29 seconds, respectively, which makes an overall difference of almost a minute.

Upon further reflection, however, I decided that timing to the second was not really critical. While timing to the second may be important in the case where you are dealing with a single short text, it would not be the case when dealing with a really long text or with multiple texts. For example, a difference of one minute is not significant in a case where it took 5 hours and 15 minutes to post-edit one version, and 5 hours and 16 minutes for the second version. When dealing with a greater volume of text, the time differences will likely be more striking. For example, one text may take 5 hours and 15 minutes to post-edit, while the other will take 3 hours and 58 minutes—a time difference worth paying attention to. It is only because the texts for the pilot testing were so short that the numbers (of minutes) seem to be so close together that it would seem to require timing to the second to differentiate between them.

To illustrate this point, assume that a text is 350 words long and that version 1 took 19 minutes to post-edit while version 2 took 20 minutes. Now imagine that the text was 10 times as long (i.e. 3,500words). At the same post-editing speed, version 1 of the longer text would take 190 minutes (3h10min) to post-edit, while version 2 would take 200 minutes (3h20min). If the text were 50 times as long (i.e. 17,500 words), then version 1 would take 950 minutes (15h50min) and version 2 would take 1,000 minutes (16h40min). In a context where a translator (or translation agency) translates hundreds of

thousands or even millions of words per year, this small difference of one minute for post-editing a 350-word text could add up to a huge difference.

Table 4 shows how small differences in total post-editing time for short texts could potentially add up to big time savings when the actual volume of translation is taken into consideration. For the purposes of this example, it was assumed that a translator's output averaged 2,000 words per day, and that he or she can edit 19 words per minute with System 1 (i.e. it would take 19 minutes to post-edit a 354-word text) and 21 words per minute with System 2 (i.e. it would take only 17 minutes to post-edit a 354-word text). Over the course of a year, the translator could save over 43 hours using System 2, although it must be noted that the whole process depends on more than just time. For instance, text type would also affect the results. Therefore, during a comparative evaluation, users should be encouraged to try the two systems using a range of texts that are typical of their usual translation fare.

		Day	Week (assuming 5 working days per week)	Month (assuming 23 working days per month)	Year (assuming 260 working days per year)
System 1	Words translated	2,000	10,000	46,000	520,000
	Post-editing time (rate = 19 words per minute)	105 minutes (1h45min)	526 minutes (8h46min)	2,421 minutes (40h21min)	27,368 minutes (456h8min)
System 2	Words translated	2,000	10,000	46,000	520,000
	Post-editing time (rate = 21 words per minute)	95 minutes (1h35min)	476 minutes (7h56min)	2,190 minutes (36h30min)	24,762 minutes (412h42min)
Time gained by using System 2		10 minutes	50 minutes	231 minutes (3h51min)	2,606 minutes (43h26min)

Table 4 – Accumulated time savings using System 2

Based on the calculations in table 4, I decided that measuring time to the nearest second is not necessary. Accordingly, I continued to measure post-editing time to the nearest minute.

For the pilot test, the raw output of both MT systems was post-edited by a single person (the researcher). Each translated text (i.e. the raw output) was divided into two parts, and the post-editor edited the first part of text A, followed by the first part of text B. The post-editor then edited the second part of text B, followed by the second part of text A. This procedure was followed for fairness as it would normally take longer to post-edit the first text (text A) than the second (text B) because the content is less familiar.

However, this approach is not necessary when the testing is extended to a pool of translators/post-editors. Half the testers were asked to post-edit the output from the first MT system (MT 1) first and the other half were asked to post-edit the output from the second MT system (MT 2) first to reduce this potential bias.

3.6.5 Ease of application

Overall I found the methodology easy to apply. The major difficulty was in deciding on the criteria and how to measure them. Once the decisions were made, applying the methodology was straightforward. However, it is worth noting that since I developed the system, I knew how it was supposed to be applied. Accordingly, I felt that it would be interesting to see how easy it would be for others to apply, and this was something I watched for in the next stage of testing.

The same situation applies for post-editing. It was easy for me to apply the methodology and restrict myself to making only non-stylistic changes. However, based on the literature on post-editing, one of the major difficulties translators face is restricting themselves to only necessary changes (i.e. no changes for style). It appears that the tendency to rewrite texts can be overcome with time, practice and a willingness to do so. I felt it would be interesting to see how others would deal with this restriction, but for the time being, I did not implement any changes.

3.6.6 Genericness of the system

It was equally easy to apply the methodology to both MT systems, which is an indication that the methodology is not system-dependent. Although it may be worthwhile adding additional MT systems into the mix to further test genericness of the evaluation system in a future stage of testing as it would be relatively straightforward to do so in principle, it would necessitate increasing the pool of post-editors, which is not feasible for the scope of this thesis. For the time being, however, I have not made any changes in how the methodology is applied.

Another way to consider genericness is how well the methodology can be applied to different texts. Since this is only a pilot test that used one text, I cannot comment on how well (generic) this methodology will work with other texts and text types, but I will be using different texts to test the refined methodology in the next stage.

3.7 Conclusion

The objective of this thesis is to develop and test an evaluation methodology that is comparative and declarative. Based on the pilot study, I feel that this objective has been fulfilled. The system I developed made it easy for me to compare the two MT systems and declare a "winner." However, the scope of the pilot study is very limited, so I will test the methodology more extensively (Chapter 4) after making some refinements.

In the extended testing, I asked four translator-evaluators to apply the refined methodology to four different source texts and their translations using the same two MT systems. In other words, each of the testers was asked to apply the methodology to one text and its two translations. These texts were chosen from other specialized fields to see

how well this methodology extends to other subject fields. However, the length of the texts chosen did not exceed 500 words as longer texts would make it more cumbersome for scorers and not likely change the overall results.

For the purposes of the extended testing, I followed the methodology used in the pilot test with the exception that I changed the order of scoring for raw output so that intelligibility was scored first, then coherence and lastly accuracy. In addition, the mechanics of the post-editing process were changed slightly as there is no need to divide the texts in two when dealing with a pool of four testers. Two (50%) of the testers were asked to post-edit the output from MT System 1 first, and the other two were asked to post-edit the output from MT System 2 first rather than splitting each version into two as for the pilot study. Thus, I was able to see if other testers could follow the methodology, and whether it allowed them to declare a "winner" (it does not matter which system actually wins).

Chapter 4 Refined Design and Larger-scale Implementation

As previously stated the objective of this thesis is to develop and test an evaluation methodology that is comparative and declarative. The results of the pilot test described in Chapter 3 suggest that the initial methodology shows promise but could benefit from some refinements. In addition, since the scope of the pilot study was limited to one tester (the developer) and one text, more extensive testing was needed in order to gain a broader and more objective overview of the value of this methodology. Consequently, as described in the following sections, a larger scale implementation of the refined methodology was carried out.

4.1 MT system and text selection

For the purposes of the larger scale testing, neither the MT system selection nor the text selection process was changed. Accordingly, the same two systems were chosen as explained in section 3.1. The two MT systems (Systran Classic (1999 version) and ReversoPro 98) were anonymized for the larger-scale testing to avoid any bias on the part of the testers. In other words, they were simply referred to as MT System 1 and MT System 2.

In a process similar to that used in the pilot test, the two MT systems were used to produce translations of four specialized (i.e. not technical) texts. Standard French as used in Canada is the source language of the texts chosen to be translated and the target language is standard Canadian English. Factors such as formatting (including fonts) are not used as criteria for the purposes of this evaluation system.

The texts chosen were approximately 350 words in length (see Appendix A). They were taken from four different Government of Canada Web sites and dealt with four different specialized topics as follows:

National Research Council – corporate overview,

Health Canada – health care,

Canada Customs and Revenue Agency – criteria for filing a tax return,

Canadian Dairy Commission – a brief history of the organization.

The English versions of these texts (see Appendix B), which were produced by human translators, were also retained from the Web sites and used (along with the source texts) to help evaluate the accuracy of the two sets of raw output from the MT systems.

The raw output from the two MT systems appears in Appendix C. Note that, as was the case with the pilot study, the source texts were edited to remove typos before being run through the MT systems so that no errors in the output resulted from input errors. The texts were not, however, "pre-edited" for anything beyond typographical errors. In other words, the texts were not altered to remove any lexical or structural ambiguities that they might contain.

4.2 Participant selection

For the purposes of the larger scale testing, four participants were asked to apply the methodology to one of the four texts selected (see section 4.1). Since this methodology is designed for translators, the main criteria for selection were education and experience in translation. All four testers have a translation degree and experience as translators. Two are working as full-time translators for the Translation Bureau, one is a

freelance translator currently completing her MA in translation and the fourth has been working in the Terminology Directorate at the Translation Bureau and is also completing her MA in translation. Moreover, all of the testers had previously been exposed to machine translation systems during the course of their studies. The testers did not reveal which text they evaluated, again to reduce any potential or perceived bias[27].

I feel that it is reasonable to apply the refined methodology to a scaled-up test using more (4) testers and texts, but using only 2 MT systems, since adding other systems would require a proportionate increase of testers, resulting in the generation of a larger amount of data than is feasible to collect and analyze within the scope of this thesis.

4.3 Refinements to methodology

For the scaled-up testing, I followed the same basic methodology as that used for the pilot test (section 2.4), but two refinements were made based on the discussions in section 3.6:

1. The order of scoring for raw output was changed so that intelligibility is scored first, then coherence and lastly accuracy.

2. Two (50%) of the testers post-edited the output from MT System 1 first, and the other two post-edited the output from MT System 2 first, rather than splitting the output from each MT system in half as was done for the pilot study.

[27] If the person running the test assigns specific texts to specific evaluators, it could be perceived as a bias in that this person could hand-pick texts in order to achieve desired results.

4.4 Testing procedure

For the purposes of the scaled-up testing, four testers met with the researcher in the computer lab at the School of Translation and Interpretation (STI). The testers were told that they were being asked to evaluate and revise (post-edit) the output of two MT systems as part of the testing of an evaluation methodology developed for my MA thesis. They were also informed that the purpose of the evaluation methodology is not to determine the absolute quality of the MT output, but rather to see if it is possible for them to determine which of the two MT systems being tested produces better quality output (i.e. is the "winner") for their purposes, as reflected by the scoring of the raw output and the time taken to post-edit it.

Each tester was asked to apply the methodology to a different text (i.e. there were four texts in total as described in section 4.1 above). The testing was carried out in two parts, and testers were given a set of instructions to help them with their task (see Appendix F). In addition, a selection of reference material (e.g. dictionaries, grammar books, Internet access) was provided for the testers to use during the evaluation and post-editing processes.

In the first part, the four testers evaluated the raw output of the two MT systems for intelligibility, coherence, and accuracy, in that order. To facilitate the scoring process, the testers were given three copies of their respective texts: texts to be scored for intelligibility and accuracy had been pre-segmented into sentences and those to be scored for coherence had been pre-segmented into paragraphs. Following each of the segments was a place for the testers to assign a score (see Appendix D).

In the second part of the testing, the testers were asked to post-edit the two translations of the text. Testers 1 and 2 were asked to post-edit the output of MT System 1 first, followed by the output of MT System 2. In contrast, Testers 3 and 4 were asked to edit the output of MT System 2 first, followed by the output of MT System1. This procedure was followed to reduce bias due to the fact that it will most likely take longer to post-edit the first text regardless of which MT system produced it. The post-editors kept track of the time taken to the nearest minute to produce a client-ready version of each text and noted it at the bottom of the documents in the space indicated.

The testers were instructed to apply Canadian spelling and usage (including typographical conventions) and to change only what was necessary to make the text accurate, intelligible and coherent. They were instructed not to make changes or revisions for style or to reflect personal preferences.

Following the application of the methodology, each tester was asked to fill out a questionnaire relating to the ease of application and effectiveness of the evaluation methodology. The completed questionnaires can be found in Appendix G.

4.5 Data analysis and evaluation

4.5.1 Raw output scores

A detailed indication of the scores received for each sentence/paragraph/text contained in the each of the target texts is presented in Appendix D, while the total scores for each of the raw output criteria are summarized in Table 5. As was the case for the pilot test, a lower score reflects a better quality output.

Criteria:	**MT System 1 (Systran)**		**MT System 2 (ReversoPro)**	
	Score	Normalized score (= score / total # of units)	Score	Normalized score (= score / total # of units)
Intelligibility:				
Tester 1	37	2.47	45	3.00
Tester 2	61	2.77	55	2.50
Tester 3	35	2.19	34	2.13
Tester 4	58	2.32	52	2.08
Total: intelligibility	191	9.75	186	9.71
Coherence:				
Tester 1	15	2.50	23	3.83
Tester 2	15	3.00	13	2.60
Tester 3	8	1.60	10	2.00
Tester 4	25	2.27	24	2.18
Total: coherence	63	9.37	70	10.61
Accuracy:				
Tester 1	31	2.07	35	2.33
Tester 2	62	2.82	58	2.64
Tester 3	37	2.31	38	2.38
Tester 4	58	2.32	53	2.12
Total: accuracy	188	9.52	184	9.47
Totals				
Tester 1	83	7.04	103	9.16
Tester 2	138	8.59	126	7.74
Tester 3	80	6.10	82	6.51
Tester 4	141	6.91	129	6.38
Totals	442	28.64	440	29.79

Table 5 – Scoring results for raw output

4.5.1.1 Discussion of group raw output scores

As a group, the testers preferred MT System 1 overall (28.64 total normalized score versus 29.79[28]). Interestingly, although MT System 1 was the preferred system overall, it actually received a higher score for intelligibility and accuracy, which indicates the output quality was poorer in these categories. However, the overall difference between scores for the two systems in these two categories was relatively small: 9.75 versus 9.71 in favour of MT System 2 for intelligibility and 9.52 versus 9.47 in favour of MT System 2 for accuracy. In contrast, however, MT System 1 outperformed MT System 2 in the category of coherence by a more significant margin: 9.37 versus 10.61 in favour of MT System 1.

These results raise some interesting points. For example, if coherence had not been used as one of the criteria, MT System 2 would be the winner (with a score of 19.18 versus 19.27). As most other evaluation systems do not test for coherence (see section 1.2.4), the fact that coherence has been included has led to the results of this test being the opposite of the results that might be obtained using other evaluation systems. However, upon closer examination of the scores assigned by individual testers, Tester 1 seems to be an "outlier"[29] in the coherence category, and has skewed the results in favour of MT System 1. The presence of outliers is very common when dealing with real data, but it does seem to indicate the need for a neutralizing mechanism when the test is on a

[28] It is important to note that the scores for each of the three scoring categories need to be normalized before the data is interpreted. If they are not, it may appear that coherence is not as important as intelligibility or accuracy due to the fact that the coherence scale is applied fewer times to a text. For example, a text comprised of 15 sentences divided into three paragraphs will have the accuracy and intelligibility scales applied 15 times each, whereas the coherence scale will be applied only four times (once for each paragraph and once for the text as a whole). As a consequence, the cumulative scores for intelligibility and accuracy will be much larger than the scores for coherence.

[29] An outlier is a data point that is located far from the rest of the data. In the case of testers 2, 3, 4, the coherence scores for the two MT systems were within two points (cumulative scores) of each other, whereas tester 1 had a spread of 8 points (cumulative score) between the two systems.

scale (i.e. four testers) where one such outlier may unduly affect the results. One possible mechanism could be to eliminate the high and low score for each of the criteria. This also points out the need for testing on a very large scale, where the effects of such outliers may not be felt as acutely.

Furthermore, in two of three categories (accuracy and intelligibility), System 1 performed more poorly, but still came out as the overall winner. Given that all three categories are considered to be equally important (see section 3.6.1), it does not seem fair that a relatively good performance in one single category (in this case, coherence) should be overridden by a poorer performance in the other two. However, as noted in the previous paragraph, Tester 1 seems to be an outlier in the coherence category, skewing the results against MT System 2. All the other testers only noted a slight difference between the two systems (i.e. differences ranging from .41 to .85 normalized points between the two systems), whereas Tester 1 had a difference of 2.12 normalized points, largely due to her coherence score (i.e. a difference of 1.33 points in favour of MT System 1). Again, this demonstrates the need to neutralize the effects of outliers in the case of a relatively small-scale test and the need to test on a very large scale in order to obtain more conclusive results.

4.5.1.2 Discussion of individual raw output scores

The individual breakdown of the results for raw output scoring is as follows:

- Tester 1 preferred MT System 1 overall (7.04 for MT System 1 versus 9.16 for MT System 2). This result was consistent for each of the three

criteria. Tester 1 declared that she preferred MT System 1 in her questionnaire.

- Tester 2 preferred MT System 2 overall (7.74 versus 8.59) as well as for each of the three criteria. This coincides with her stated preference on the completed questionnaire.
- Tester 3 preferred MT System 1 overall (6.10 to 6.51) and for two of the three categories (intelligibility was the exception but only by a very small margin: 2.13 versus 2.19).
- Tester 4 preferred MT System 2 to MT System 1 (6.38 versus 6.91) as well as for each three of the scoring categories.

Thus, two of the testers (1 and 3) preferred MT System 1, and two (2 and 4) preferred MT System 2, although for the group as a whole, MT System 1 was the "winner." Three of the four testers gave winning scores that corresponded with the overall results to all three of the criteria: only tester 3 gave a non-winning score to one of the categories.

The fact that some of the testers preferred MT System 1 while others preferred MT System 2 seems to provide evidence that this evaluation methodology really is generic: it is not biased to one system or the other but allows users to compare both and select the one that appears to best meet their individual needs. If all the testers had consistently chosen the same system, it would be impossible to tell whether that was simply the better system for everyone, or whether the proposed evaluation methodology somehow favoured that particular system. However, the methodology will need to be

tried out on a much wider range of systems before any definitive claims can be made with regard to its genericness (see section 5.2.2).

4.5.2 Post-editing results

The post-edited texts and total post-editing time taken for each appear in Appendix E. The details of time taken by each tester and average number of words post-edited per minute appear in Table 6. Whichever text is post-edited first will inevitably take longer; accordingly, Testers 1 and 2 post-edited MT System 1 output first, while Testers 3 and 4 post-edited MT System 2 output first in order to reduce this bias.

	MT System 1 (Systran)			**MT System 2 (Reverso Pro)**		
	Total time in minutes	Number of words in raw output	Normalized time (= total time / total # of words in raw output) in words per minute (wpm)	Total time in minutes	Number of words in raw output	Normalized Time (= total time / total # of words in raw output) in words per minute (wpm)
Tester 1	21 min	341	16 wpm	15	333	22 wpm
Tester 2	15 min	368	25 wpm	10	392	40 wpm
Tester 3	13 min	359	27 wpm	17	368	22 wpm
Tester 4	15 min	400	27 wpm	25	403	16 wpm
Total	64 min	1,468		67 min	1,496	
Average	16 min		23 wpm	16 min 45 sec		22 wpm

Table 6 – Post-editing times

4.5.2.1 Discussion of group post-editing times

Overall, based on total post-editing effort (reflected by time taken to post-edit), the group of testers preferred MT System 1.[30] This result corresponds with the results for raw output. Accordingly, for the pool of testers, the results indicate there is a link

[30] Although the margin of preference appears slim—16 minutes for MT System 1 versus 16 minutes 45 seconds for MT System 2 – it is important to remember, as discussed in section 3.6.4, that a small difference on a short text could add up to a significant difference over the course of a year's worth of translations.

between the quality of the raw output and the amount of post-editing required (as reflected by time taken).

It is important to note that whichever text was post-edited first by any given post-editor took longer and thereby had poorer results. However, when the results for the group as a whole are considered, the bias has been reduced (see section 3.6.4) and the results are consistent with the scores for raw output.

4.5.2.2 Discussion of individual post-editing times

It is not possible to examine the individual results in order to establish whether or not there is a relationship between the quality of the raw output and post-editing times because, as explained in section 4.1, the bias associated with post-editing one text after another was reduced for the group as a whole by having testers 1 and 2 post-edit the output from MT System 1 first, while testers 3 and 4 post-edited the output from MT System 2 first. However, this did not serve to reduce the bias on an individual basis, only for the group as a whole.

Testers 1 and 2 post-edited the output from MT System 1 first and in both cases, it took longer, as expected. Tester 1 preferred MT System 1 and stated that it was easier to post-edit (i.e. it took less time) the second text (MT System 2) although the second text required a greater amount of editing. Tester 2 preferred MT System 2 overall, which was reflected in the scores for both raw output and post-editing time.

Testers 3 and 4 post-edited the output from MT System 2 first and, again, in both cases, it took longer. Tester 3 preferred MT System 1 overall, which was reflected in the

scores for raw output and post-editing time. Tester 4 preferred MT System 2, which was reflected in the raw output scores.

It is interesting to note that the individual translators did not automatically prefer the system that was faster to post-edit (i.e. the second one they used). In both cases, (i.e. where MT System 1 was post-edited first, and where MT System 2 was post-edited first), there was one tester that preferred the first system and one that preferred the second. This raises the question as to whether it is preferable from a translator's point of view (e.g. for job satisfaction) to do a larger amount of easy editing, or a smaller amount of more challenging editing. From a purely financial perspective, the better system is likely to be the one that allows the post-editing (regardless of whether it is of the type "more but easy" or "less but challenging") to be carried out in the shortest amount of time. However, the results of this test would seem to indicate that speed of post-editing was not the only factor the testers took into account when making their decision about which system they preferred. An interesting topic for further research could be to try to determine what other factors come into play when selecting an MT system for use (see section 5.2.4).

4.6 Suggestions for further refinements to the methodology

The development of a good methodology is a cyclical process. Consequently, based on the results of this larger-scale testing, I would propose some further refinements to the evaluation methodology. These additional refinements are derived from two main sources: 1) my own observations, and 2) feedback received from each of the four testers in the form of a questionnaire that was filled out after completing the evaluation methodology (see Appendix G).

4.6.1 Weighting of criteria in scoring system

As discussed in sections 3.3 and 3.61, the three criteria of intelligibility, coherence and accuracy were all assumed to be equally important to the quality of the raw output. In fact, when considered individually, the results of three of the four testers as well as the results of the pilot study show that there is a correlation between the scores for the three criteria. Only Tester 3 had one score (for intelligibility) that was not consistent with her overall results.

Although it is possible that an outlier may skew the collective results for a relatively small-scale test (i.e. four testers) as discussed in section 4.5.1.1, it is not likely that the scores assigned by an outlier would skew the results of a very large-scale test to the same degree.

However, although the three criteria of intelligibility, coherence and accuracy are meant to be equally important, upon further reflection I realized that it is possible that this may not turn out to be the case if the methodology is applied as originally devised. In the original methodology, accuracy is measured using a 4-point scale, while intelligibility and coherence are measured with a 5-point scale. This difference in the scoring scales could inadvertently result in one of the metrics having more weight (or being more important) than the others. The fact that a completely inaccurate sentence would receive a score of 4, whereas a completely unintelligible sentence would receive a score of 5 means that a system with good accuracy but poor intelligibility would automatically perform more poorly overall than a system that has good intelligibility but poor accuracy. For example, if MT System 1 always got the best possible marks for intelligibility (i.e. 1 out

of 5) but the worst marks for accuracy (i.e. 4 out of 4), a 5-sentence text would receive a score of 25 (5 points for each of the 5 sentences as follows: 1 for intelligibility and 4 for accuracy). However, if that same text were run through MT system 2, which got the best marks for accuracy (i.e. 1 out of 4), but the worst marks for intelligibility (i.e. 5 out of 5), it would receive a score of 30 (6 points for each of the 5 sentences as follows: 1 for accuracy and 5 for intelligibility). As a result, MT system 1 would be considered the better system as 25 is lower than 30. However, if both scales were equal (i.e. if both had 5-point scales), then the two MT systems would be equally good or bad, which seems more fair as accuracy and intelligibility are considered to be equally important to the quality of the raw output.

Accordingly, it would seem reasonable to refine the scoring metrics so that all three have the same point scale: that is, change the scoring scale for accuracy to a 5-point scale so that all three of the raw output metrics would be of equal importance.

4.6.2 Level of detail in scoring system

No refinement seems to be required to make the scoring system more fine-grained[31] (see discussion in section 3.6.2). None of the testers indicated that they felt the scales should be finer grained, although this question was not specifically asked. However, one tester (tester 2) indicated that she experienced some problem when trying to decide on the seriousness of errors and another (tester 4) expressed doubt about the

[31] Although, as noted in section 4.6.1, it may be desirable to increase the accuracy scale from 4 points to 5 points, not to make it finer-grained per se, but rather to make it equal in weight to the 5-point scales used to measure intelligibility and coherence.

need for differentiating between the seriousness of errors and suggested that simply keeping track of the number of errors might be sufficient for scoring purposes.

This latter comment raises an interesting point. It is possible that it could take the same amount of time to fix minor errors as it does serious errors. This premise may be worth testing in future research (see section 5.2.1). For example, two scoring systems could be devised—one that resembles the scale used in this thesis where errors are graded according to seriousness, and the other which simply keeps track of the total number of errors. Testers could be asked to apply them both and compare the results with post-editing effort. If it did turn out to be the case that simply keeping track of the total number of errors was sufficient, this would make the methodology simpler to apply as users would not need to evaluate the seriousness of an error, they could simply note that an error had occurred.

4.6.3 Order of application of raw output criteria

The order of application of raw output criteria—intelligibility followed by coherence followed by accuracy—does not seem to require any further refinements.

4.6.4 Post-editing time measurement

All four testers found the scoring metric (total time taken) for post-editing easy to apply. In addition, as noted in section 4.3.2.1, the results for the group as a whole correspond to the results for raw output. In all cases, it took longer to post-edit the first text, regardless of quality of raw output. The first text likely takes longer, not only

because its content is less familiar, but also because the research time required to find the correct terminology is included in the post-editing time for the first text, but this terminological research does not need to be repeated when post-editing the second text (i.e. the terminology is same for both sets of raw output, so that the first text is unfairly charged with research time for both texts). One of the testers (Tester 2) indicated that finding the terminology was the biggest problem in post-editing.[32] One way of adjusting for this could be to subtract the term research time from the post-editing time for the first text. However, no refinement to the methodology seems to be necessary because adjusting the order in which texts (or parts of texts) are post-edited seems to reduce the overall bias.

4.6.5 Ease of application

All four of the testers found that the methodology (both the post-editing metrics and the raw output scoring) was relatively straightforward to apply. One tester (Tester 4) suggested that the instructions for scoring raw output include a clear definition of each category (i.e. intelligibility, coherence, and accuracy). This refinement would be easy to accomplish, and because all users would likely benefit from such clarification, I would recommend adding definitions to the instruction sheet as an additional refinement of this methodology. Furthermore, adding more details regarding the target audience and the

[32] Note that while this was the case for the testers in our experiment, who were not necessarily specialists in the subject matter contained in the texts they were asked to post-edit, this may not be the case for all testers. Presumably, translators who are evaluating multiple MT systems with a view to purchasing one of them would test all the systems using "real" texts of the type that they normally translate as part of their regular workload.

required level of quality to the post-editing instructions would also probably be of benefit to testers.

Moreover, I would also add examples of coherence, as suggested by Tester 3, who found the coherence metric more difficult to apply than the other two. This refinement would be more cumbersome to implement, as coherence is scored at paragraph or text level, so that the examples would need to be longer than just a few words.

In addition, the impact of formatting (e.g. tables and bullet lists) could also be taken into consideration for coherence. Further testing on texts containing a more extensive range of formatting would need to be carried out.

Finally, one tester (Tester 3) indicated that the methodology is useful from a tester's perspective, but as a potential purchaser she would also need to take into account the cost and the effort required to carry out the methodology. In my opinion, the methodology can be implemented at little or no cost by using "demo" versions of MT software for the testing. It is also possible to contact manufacturers and ask for an evaluation copy, which can be returned if the potential purchaser decides not to go ahead with the purchase. Most manufacturers are very willing to provide an evaluation copy as they hope to make a sale from it. As well, there is not a great deal of additional effort involved: potential purchasers would be able to test the systems by using actual texts that they have translated or are likely to translate and pre-segmenting them into paragraphs and sentences. It seems reasonable to assume that if a translator is thinking of investing in any type of technology, then that translator will be prepared to invest some time in evaluating the available options and choosing the best one rather than simply buying the

first one encountered. The evaluation methodology outlined in this thesis facilitates, rather than hinders, such an evaluation process.

4.6.6 Genericness of the system

Testers experienced no great difficulties applying the evaluation methodology to either of the two MT systems. This would seem to indicate that the methodology could be applied to any MT system, though additional testing would need to be carried out on a wider range of systems in order to verify this with any certainty. The evaluation system is generic due in large part to the manner in which texts are set up: that is, they are pre-segmented into scoring units as discussed in section 2.3, with a space for the tester to write the score as well as an indication of number of points in the scoring scale. This means that it would be easy to apply to any MT system. Indeed, the fact that different testers chose different systems, as mentioned in section 5.4.1.2, is a good indication that this methodology is generic.

4.7 Summary

All four of the testers indicated that they found the methodology to be useful. In addition, the results of the scaled-up testing confirmed the results of the pilot study: application of the evaluation methodology to the output from two MT systems allows the evaluator to compare the results and declare which system he or she thinks is better. Furthermore, when the results of a pool of evaluators are considered as a whole, the methodology allows one MT system to be declared better than the other. However, a few further refinements to the methodology, in particular changing the scoring metric for

accuracy to a 5-point scale as discussed in 4.6.1 and providing some examples for the coherence metric and definitions of intelligibility, coherence and accuracy would make it more effective.

Chapter 5 Conclusion

5.1 General evaluation of the methodology

As stated in section 0.2, the objectives of this thesis were to *develop* and *test* an evaluation methodology for assessing the quality of MT output. These two objectives have been met: the evaluation methodology that was developed was presented in chapters 2 and 3, and the results of the testing were presented in chapter 4.

The goal was to produce an evaluation methodology that met the following criteria: generic, comparative, task-based and declarative. The methodology is generic in that it was successfully applied to two different MT systems[33]. It is both comparative and declarative in the sense that evaluators can compare the output and declare which system produces better output (for their purposes). As well, it is task-based in the sense that a specific user group (translator-users working on a freelance basis or for a small translation company) and the purpose for which it requires the translation (informative text for delivery to a client) were selected as the basis for evaluation.

Based on the results of both the pilot study and the scaled-up testing, it appears that generic, comparable, declarative, task-based evaluation of MT systems can be effectively carried out using the metrics developed in this thesis. Furthermore, these metrics are reusable by other translator-users, as was demonstrated in the scaled-up testing (chapter 4). All four of the testers were able to apply the methodology and declare which system they preferred.

[33] As noted in section 5.2.2, the extent of the genericness can only be confirmed through further testing.

In addition, it appears that using total time as a measure of post-editing effort is a valid metric for evaluating the quality of the output—confirming the results for the scoring of the raw output—assuming the time bias is neutralized. This bias can be neutralized by having individual translators divide the output from each MT system into two parts and alternating the post-editing (see section 2.4.2). By the same token, this bias needs to be neutralized when testing in a larger group, which can be done by having 50% of the testers post-edit the raw output from MT System 1 first, and the other 50% post-edit the raw output from MT System 2 first (see section 3.6).

5.2 Suggestions for future research

This thesis represents a preliminary exploration of the issues involved in generic, comparable, task-based, declarative evaluation of MT output. Further work that could be carried out includes additional refinements to the methodology, additional types of testing, MT error analysis, and integration of the methodology proposed here into a more comprehensive methodology.

5.2.1 Additional refinements to the proposed methodology

As discussed in section 4.6, the development of a good methodology is a cyclical process and further refinements are always possible. Potential additional refinements to the evaluation methodology include changing the scoring metric for accuracy to a 5-point scale as discussed in 4.6.1, investigating whether the seriousness or the number of errors has more impact on the results (see section 4.6.2), as well as adding clear definitions for

the scoring metrics, more detailed instructions for post-editing and examples for scoring coherence as discussed in section 4.6.5.

5.2.2 Additional types of testing

The proposed methodology needs to be tested more extensively. For example, larger-scale testing with a larger pool of translator-users to carry out evaluations (variable testing) would help validate and verify the conclusions reached in this thesis. More extensive testing using a larger corpus of texts, including more text types, lengths and areas of specialization, and using other language pairs would also help validate and verify the results. Tests evaluating additional MT systems would further test the genericness of the methodology. In addition, tests using source documents with complicated formatting such as tables, etc. would investigate how effectively various MT systems handle these features and any impact on coherence.

5.2.3 MT error analysis

The ability to edit and customize the user-defined dictionaries in MT systems to improve the MT systems' performance is an important factor when considering purchasing an MT system. As a result, the extent to which these dictionaries can be edited, and how easily, needs to be investigated. This would entail the additional task of collecting and classifying recurring errors during the post-editing stage with an eye to improving the MT system by customizing the system dictionaries and/or reporting the errors to the developers so that they may improve future versions of the MT system.

Improvability is important because the more the raw output can be improved, the greater its quality, which will (hopefully) reduce post-editing time and thus the cost of producing the translation.

5.2.4 Integration of the proposed methodology into a more comprehensive methodology

It should also be noted that other factors (e.g. user-friendliness, ease of customizing dictionaries, etc.) are also important for a global assessment of a system's usefulness, but are beyond the scope of this thesis. Another area worth investigating is the impact of incorporating MT systems into the translation process and how this affects translators' job satisfaction, including how they react to the type of post-editing required (i.e. whether they prefer "more but easy" or "less but challenging") as discussed in section 4.5.2.2.

I feel that it would be useful to take the methodology that has been developed in this thesis and integrate it into a larger more comprehensive methodology that also takes into account these other factors.

5.3 Concluding remarks

As translators continue to face the pressure of increasing volumes of translation and shorter deadlines, more of them may begin turning to technology, such as machine translation systems, to help them work more productively.

According to Allied Business Intelligence (ABI 2002: 5-21), "The MT share of the translation market is relatively small at the moment, but it has an exponential growth

potential." ABI (2002: 5-23) notes that in 2001, worldwide revenue for MT was approximately US$72.9 million. Under the ABI moderate forecast, this will rise to US$114.2 million in 2007 with a cumulative average annual growth rate of 8%, and under the aggressive forecast, this value will reach US$133.8 million with a growth rate of 11%.

However, as more translators begin to show an interest in using MT systems, a greater number of developers will likely introduce such systems to the market. Translators will therefore be faced with an increasingly wide range of products to evaluate in order to determine which one can best meet their needs. While a large company may have the resources—both human and financial—to experiment with multiple systems or to conduct extended testing, this is not a luxury afforded to most freelance translators or small translation agencies. The latter group will require a methodology that is straightforward and efficient to apply in order to help them determine which system best meets their requirements. It is hoped that the evaluation methodology described in this thesis will go some way towards meeting the needs of such translators.

Glossary

ALPAC: see **Automatic Language Processing Advisory Committee.**

Automatic Language Processing Advisory Committee (ALPAC): committee set up by the United States government in 1964 to evaluate the state of MT research. In 1966, ALPAC produced a report entitled *Language and Machines: Computers in Translation and Linguistics* (commonly referred to as the ALPAC report).

black-box evaluation: type of evaluation where the evaluator looks at MT operations uniquely in terms of input (source text) and output (target text). The evaluator does not have access to the internal workings, the underlying algorithms and source code of the MT system. See also **glass-box evaluation**.

CAT: see **computer-assisted translation**.

client-ready text: text that is of sufficient quality to deliver to a client or end-user. Typically, the translator or client decides on the level of quality required.

comparative evaluation: evaluation where the output or raw translations of the same source text from two or more MT systems are compared with each other in terms of quality, as opposed to comparing the MT output with translation(s) done by human translators.

computer-assisted translation (CAT): use of computers to aid translators during the translation process. CAT tools include word processors and electronic dictionaries, corpus analysis tools and translation memories. See also **machine translation**.

conventional post-editing: process where the raw output of an MT system is "totally" or "completely" revised so that the target text resembles an original text. Changes for style are made. See also **partial post-editing**.

criticism: process of evaluation a text (or translation) after it is finished or published. Criticism generally applies to literary or artistic translations.

declarative evaluation: evaluation measuring the ability of two or more MT systems to handle "real" texts allowing the evaluator to declare which system is better or "the winner," based on the results of the scored quality metrics.

FAHQT: see **fully automatic high-quality translation.**

fully automatic high-quality translation (FAHQT): output produced by an MT system that is of high quality requiring no human intervention of any type and is indistinguishable from translations done by a human translator.

generic evaluation: evaluation methodology used to evaluate the output of any MT system regardless of its architecture, the language-pair(s) involved or the text type. Using the same, generic methodology for evaluation will ensure that the results are comparative.

glass-box evaluation: type of evaluation where the evaluator has access to the internal workings, the underlying algorithms and source code of the MT system. See also **black-box evaluation**.

machine translation (MT): use of computers to produce draft translations without any human intervention.

maximal post-editing: see **conventional post-editing**.

MT: see **machine translation.**

outlier: data point that is located far from the rest of the data. Given a mean and standard deviation, a statistical distribution expects data points to fall within a specific range. Those that do not are called outliers and may require further investigation. In practice, nearly all experimental data samples are subject to contamination from outliers, a fact which reduces the real efficiency of theoretically optimal statistical methods.

partial post-editing: process of revising the raw output of an MT system so that the syntactic norms or standards of the target language are respected and to ensure that the meaning is accurate compared with the source text. No stylistic changes are made. See also **conventional post-editing**.

post-editing: process where raw MT output is revised in order to "clean up" the MT output so that it can be delivered to the client.

rapid post-editing: see **partial post-editing**.

raw output: target text produced by a machine translation system with no human intervention.

revision: process of controlling document production (or translation) for accuracy, completeness, stylistic appropriateness, etc.

task-based approach: approach where a specific user group and its needs and requirements are used as the defining parameters for the evaluation.

References

Allen, Jeffrey (2003) "Post-editing" in *Computers and Translation: A Translator's Guide*, Harold Somers (ed), Amsterdam: John Benjamins, 297-317.

Allied Business Intelligence report (2002) *Language Translation, Localization and Globalization: World Market Forecasts, Industry Drivers, and eSolutions.*

Automatic Language Processing Advisory Committee (ALPAC) (1966) *Language and Machines: Computers in Translation and Linguistics.* <http://www.nap.edu/books/ARC000005/html/index.html > accessed October 31, 2002.

Andrés Lange, Carmen and Winfield Scott Bennett (2000) "Combining Machine Translation with Translation Memory at Baan" in *Translating Into Success: cutting-edge strategies for going multilingual in a global age*, Robert C. Sprung (ed), Amsterdam: John Benjamins, 203-218.

Arnold, Doug (2003) "Why translation is difficult for computers" in *Computers and Translation: A Translator's Guide*, Harold Somers (ed), Amsterdam: John Benjamins, 119-142.

Arnold, Doug, Lorna Balkan, Siety Meijer, R. Lee Humphreys, and Louisa Sadler (1994) *Machine Translation, An Introductory Guide*, Manchester: NCC Blackwell < http://www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html > accessed October 20, 2003.

Austermühl, Frank (2001) *Electronic Tools for Translators*, Manchester: St. Jerome Publishing.

Bédard, Claude (2000) "Mémoire de traduction cherche traducteur de phrases..." in *Revue TRADUIRE*, issue 186 < http://www.terminotix.com/t_fr/index.htm > accessed February 1, 2004.

Bowker, Lynne (2002) *Computer-Aided Translation Technology: A Practical Introduction*, Ottawa: University of Ottawa Press.

Bowker, Lynne (2001) "Towards a Methodology for a Corpus-Based Approach to Translation Evaluation" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 345-364.

Brace, Colin (2000) "Language Automation at the European Commission" in *Translating Into Success: cutting-edge strategies for going multilingual in a global age*, Robert C. Sprung (ed), Amsterdam: John Benjamins, 219-224.

Canadian Translation Industry Sectoral Committee (1999) *Survey of the Canadian Translation Industry – Human Resources and Export Development Strategy*, GDSourcing Research & Retrieval < http://www.gdsourcing.ca/works/Translation.htm > accessed August 6, 2003.

Church, Kenneth W. and Eduard H. Hovy (1993) "Good Applications for Crummy Machine Translation" in *Machine Translation* 8, 1993 <http://www.isi.edu/natural-language/people/hovy/papers/93churchhovy.pdf> accessed November 10, 2003.

EAGLES (Expert Advisory Group on Language Engineering Standards) (1999) *EAGLES Evaluation of Language Processing Systems*, Center for Sprogteknologi: Copenhagen, Denmark <http://issco-wjww.unige.ch/projects/eagles/ewg99/ > accessed October 31, 2002.

EAGLES (Expert Advisory Group on Language Engineering Standards) (1996) *EAGLES Evaluation of Natural Language Processing Systems*, Center for Sprogteknologi: Copenhagen, Denmark <http://issco-www.unige.ch/projects/ewg96/ewg96.html > accessed October 31, 2002.

EAMT-CLAW (2003) *EAMT-CLAW'03: Controlled Language Translation (Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop)*, Dublin/Geneva: Dublin City University and EAMT.

Fernández Guerra, Ana (2000) *Machine Translation: Capabilities and Limitations*, Valencia: Universitat de València.

Fromkin, Victoria, Robert Rodman, Neil Hultin and Harry Logan (1997) *An Introduction to Language,* Toronto/Montréal: Harcourt Brace.

Gerzymisch-Arbogast, Heidrun (2001) "Equivalence Parameters and Evaluation" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 227-242.

Halliday, M.A.K and Ruqaiya Hasan (1976) *Cohesion in English*, London: Longman.

Hönig, Hans G. (1998) "Positions, Power and Practice: Functionalist Approaches and Translation Quality Assessment" in *Translation and Quality*, Christina Schäffner (ed), Clevedon/Philadelphia/Toronto/Sydney/Johannesburg: Multilingual Matters, 6-27.

Horguelin, Paul A. and Louise Brunette (1998) *Pratique de la revision*, 3rd edition, Brossard, Quebec: Linguatech.

House, Juliane (2001) "Translation Quality Assessment: Linguistic Description versus Social Evaluation" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 243-257.

House, Juliane (1997) *Translation Quality Assessment. A Model Revisited*, Tübingen: Gunter Narr.

Hutchins, W. John (1998) "Evaluation of machine translation and translation tools" in *Survey of the State of the Art in Human Language Technology*, Ronald A. Cole (ed), Oregon: Center for Spoken Language Understanding and University of Pisa. < http://cslu.cse.ogi.edu/HLTsurvey/ch13node5.html > accessed October 20, 2003.

Hutchins, W. John and Harold L. Somers (1992) *An Introduction to Machine Translation,* London: Academic Press.

International Standards Organization 9126 (1991), *ISO/IEC 9126 Information Technology "Software Product Evaluation" Quality Characteristics and Guidelines for Their Use.* Geneva.

King, Margaret (1996) "On the notion of validity and the evaluation of machine translation systems" in *Terminology, LSP and Translation, Studies in Language Engineering in Honour of Juan C. Sager*, Harold Somers (ed), Amsterdam: John Benjamins, 189-205.

Krings, Hans P. (2001) *Repairing Texts: Empirical Investigations of MT Post-editing Processes*, Kent, Ohio: Kent State University Press.

Lee-Jahnke, Hannelore (2001) "Aspects pédagogiques de l'évaluation en traduction" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 258-271.

Lewis, Derek (1992) "Computers and Translation" in *Computers and Written Texts*, Christopher S. Butler (ed), Oxford: Blackwell, 75-113.

L'Homme, Marie-Claude (1999) *Initiation à la traductique*. Brossard, Québec: Linguatech.

Lockwood, Rose (2000) "Machine Translation and Controlled Authoring at Caterpillar" in *Translating Into Success: cutting-edge strategies for going multilingual in a global age*, Robert C. Sprung (ed), Amsterdam: John Benjamins, 187-202.

Loffler-Laurian, Anne-Marie. (1996) *La traduction automatique,* Paris: Septentrion.

Macklovitch, Elliott and Graham Russell (2000) "What's Been Forgotten in Translation Memory" in *Envisioning Machine Translation in the Information Future – Proceedings of the 4th Conference of the Association for Machine Translation in the Americas AMTA 2000 Cuernavaca, Mexico, October 10-14, 2000,* Berlin/Heidelberg/New York: Springer-Verlag, 137-146.

Mariani, Joseph (ed) (1996) "Evaluation" in *Survey of the State of the Art in Human Language Technology*, Ronald A. Cole (ed), Oregon: Center for Spoken Language Understanding and University of Pisa. <http://cslu.cse.ogi.edu/HLTsurvey/> accessed November 11, 2002.

Martínez Melis, Nicole and Amparo Hurtado Albir(2001) "Assessment in Translation Studies: Research Needs" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 272-287.

Melby, Alan and C. Terry Warner (1995) *The Possibility of Language,* Amsterdam: John Benjamins.

Newmark, Peter (1991) *About Translation*, Clevedon: Multilingual Matters.

Newmark, Peter (1988) *Approaches to Translation*, New York: Prentice Hall.

Nida, Eugene A. and Charles A. Taber, (1969) *The Theory and Practice of Translation*, Leiden, Netherlands: E. J. Brill.

Nyberg, Eric, Teruka Mitamura, and Willem-Olaf Huijsen(2003) "Controlled Language for Authoring and Translation" in *Computers and Translation: A Translator's Guide*, Harold Somers (ed), Amsterdam/Philadelphia: John Benjamins, 245-281.

Papegaaij, Bart and Klaus Schubert (1988) *Text Coherence in Translation,* Dordrecht, Holland: Foris.

Rinsche, Adriane (1994) "Pour une méthodologie de l'évaluation de la TA" in *TA-TAO : Recherches de pointe et applications immédiates, Troisièmes Journées scientifiques du réseau thématique « Lexicologie, Terminologie, Traduction »* André Clas (dir), Montréal: FMA Beyrouth, 265-274.

Sager, Juan C. (1994) *Language Engineering and Translation: Consequences of Automation,* Amsterdam: John Benjamins.

Schäfer, Falko (2003) "MT Post-Editing: Experiences at SAP" in *EAMT-CLAW'03: Controlled Language Translation (Proceedings of the Joint Conference combining*

the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop), Dublin/Geneva: Dublin City University and EAMT, 185-192.

Schäffner, Christina (1998a) "From 'Good' to 'Functionally Appropriate': Assessing Translation Quality" in *Translation and Quality*, Clevedon/Philadelphia/Toronto/Sydney/Johannesburg: Multilingual Matters, 1-5.

Schäffner, Christina (ed) (1998b) *Translation and Quality*, Clevedon/Philadelphia/Toronto/Sydney/Johannesburg: Multilingual Matters.

Schwarzl, Anja (2001) *The (Im)possibilites of Machine Translation*, Frankfurt am Main/Berlin/Bern/Bruxelles/New York/Oxford/Wien: Peter Lang.

Somers, Harold (1999) "Review Article: Example-based Machine Translation," *Machine Translation*, vol. 14, no. 2, 113-157.

Trujillo, Arturo (1999) *Translation Engines: Techniques for Machine Translation*, London: Springer-Verlag.

van der Meer, Jaap (2003) "At Last Translation Automation Becomes a Reality: An Anthology of the Translation Market" in *EAMT-CLAW'03: Controlled Language Translation (Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop)*, Dublin/Geneva: Dublin City University and EAMT, 180-184.

Vanni, Michelle and Florence Reeder (2000) "How Are You Doing? A Look at MT Evaluation" in *Envisioning Machine Translation in the Information Future: 4th Conference of the Association in the Americas, AMTA 2000*, John S. White (ed), Berlin/Heidelberg/New York: Springer-Verlag, 109-116. <http://www.mitre.org/support/papers/tech_papers99_00/vanni_evaluation/index.shtml> accessed July 4, 2000.

Van Slype, Georges (1979) *Critical Study of Methods for Evaluating the Quality of Machine Translation*, Final Report, Bureau Marcel van Dijk / European Commission, Brussels. < http://issco-www.unige.ch/projects/isle/van-slype.pdf > accessed September 2, 2003.

Waddington, Christopher (2001) "Different Methods of Evaluating Student Translations: The Question of Validity" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 311-325.

White, John S. (2003) "How to evaluate machine translation" in C*omputers and Translation: A Translator's Guide*, Harold Somers (ed), Amsterdam: John Benjamins, 211-244.

White, John S. (ed.) (2000a) *Envisioning Machine Translation in the Information Future (Proceedings of the 4th Conference of the Association for Machine Translation in the Americas)*, Berlin/Heidelberg/New York: Springer-Verlag.

White, John S. (2000b) "Contemplating Automatic MT Evaluation" in *Envisioning Machine Translation in the Information Future: 4th Conference of the Association in the Americas, AMTA 2000*, John S. White (ed), Berlin/Heidelberg/New York: Springer-Verlag, 100-108.

Williams, Malcolm (2001) "The Application of Argumentation Theory to Translation Quality Assessment" in *META, Evaluation: Parameters, Methods, Pedagogical Aspects*, H. Lee-Jahnke (dir), Vol 46, no 2, Montréal: Les Presses de l'Université de Montréal, 326-344.

Word, Nigel (2000) "Machine Translation" in *Speech and Language Processing, An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Daniel Jurafsky and James H. Martin (eds), Upper Saddle River, New Jersey: Prentice-Hall, 799-825.

Zaenan, Annie (ed) (1996) "Multilinguality" in *Survey of the State of the Art in Human Language Technology*, Ronald A. Cole (ed), Oregon: Center for Spoken Language Understanding and University of Pisa. <http://cslu.cse.ogi.edu/HLTsurvey/> accessed November 11, 2002.

Appendix A Source Texts

1. SARS text (used for pilot study in Chapter 3)

Source text: from Santé Canada En Direct

<http://www.hc-sc.gc.ca/francais/protection/mises_garde/sras/questions6.html>

accessed April 7, 2003

354 words

Questions et réponses - Syndrome respiratoire aigu sévère (SRAS)

Santé au travail

Q. Comment Santé Canada appuie-t-il les fonctionnaires fédéraux en ce qui a trait au SRAS?

R. Le rôle du Programme de santé au travail et de sécurité du public (PSTSP) de Santé Canada consiste à conseiller les fonctionnaires fédéraux à propos des questions d'hygiène professionnelle et de sécurité. Les professionnels de la santé du PSTSP de Santé Canada étaient sur place aux aéroports Pearson, de Vancouver et de Dorval pour organiser des séances d'information à l'intention des fonctionnaires fédéraux. Ils continuent d'organiser des séances d'information pour les employeurs et les employés en milieux de travail partout au pays, et de répondre aux demandes de renseignements quotidiennes de la part des employeurs.

Le Programme de santé au travail et de sécurité du public a émis des avis, et des membres de son personnel se sont entretenus et réunis avec des représentants d'un certain nombre de ministères fédéraux en vue de leur fournir de l'information et des conseils sur les mesures de santé et de sécurité jugées appropriées pour leurs employés. Les représentants du PSTSP ont assisté à une vidéoconférence en présence des employés travaillant dans sept missions à l'étranger situées dans les régions touchées afin de leur fournir de l'information sur la protection des voies respiratoires et de discuter d'autres préoccupations de santé générale au sujet de SRAS.

Q. Pourquoi Santé Canada ne recommande-t-il pas l'usage de masques par les fonctionnaires fédéraux tels que les agents de douanes aux aéroports Pearson et de Vancouver?

R. Le rôle du PSTSP de Santé Canada consiste à fournir des conseils aux fonctionnaires fédéraux sur des questions d'hygiène professionnelle et de sécurité.

Santé Canada croit que les employés du gouvernement fédéral ne risquent pas de devenir malades, car les contacts des employés avec les passagers ou les clients sont d'une durée relativement courte et ne sont pas considérés comme étroits. Par conséquent, Santé Canada considère que l'usage des masques n'est pas nécessaire.

Les employeurs peuvent utiliser cette information et ces conseils afin de déterminer les mesures de santé et de sécurité qu'ils jugent appropriées pour leurs employés.

2. Canadian Dairy Commission text (used by Tester 1 for scaled-up testing in Chapter 4)

Source text: from Canadian Dairy Commission site

< http://www.cdc.ca/cdc/main_f.asp?catid=663&page=1761 >

accessed March 13, 2004

339 words

Historique

Depuis la nomination du premier Commissaire du lait du Dominion en 1890, le gouvernement fédéral a joué un rôle actif dans la mise au point et l'application des politiques et programmes afin d'appuyer l'industrie laitière.

Premiers programmes

Les premiers programmes comprenaient un service de wagons réfrigérés pour le beurre en 1895, des salles fraîches pour la maturation du fromage en 1902, des tests pour les vaches laitières en 1902 et le classement du beurre et du fromage destinés à l'exportation en 1923. Plus tard, en 1935, le gouvernement du Canada adoptait une subvention provisoire pour le fromage et le beurre. Au cours des années quarante et cinquante, il a lancé des programmes de soutien des prix, d'exportation de produits excédentaires et de limitation des importations, programmes qui ont été continués après la création de l'Office de stabilisation des prix agricoles en 1958.

Bien qu'il ait offert la structure voulue pour les activités de stabilisation des prix, l'Office de stabilisation des prix agricoles n'était pas en mesure de s'attaquer à deux problèmes majeurs qui affectaient sérieusement le secteur laitier :

- 1) le manque de coordination entre les politiques fédérales et provinciales, et
- 2) l'absence d'un mécanisme efficace de contrôle de la production laitière.

Un tel mécanisme s'avérait nécessaire pour assurer des prix de soutien raisonnables pour les produits laitiers à entreposer afin de renforcer les revenus des producteurs. De nouveaux mécanismes étaient aussi requis pour contrôler les surplus de production coûteux. En 1963, le gouvernement fédéral a convoqué une Conférence canadienne de l'industrie laitière, ce qui a mené à la création du Comité consultatif de l'industrie laitière, la même année. Le rapport final de ce Comité, présenté en 1965, recommandait la creation de la Commission canadienne du lait.

La Commission canadienne du lait (CCL) a été établie par la proclamation de la Loi sur la Commission canadienne du lait, le 31 octobre 1966. Par la création de cette société de la Couronne particulière, le gouvernement fédéral confirmait son engagement d'assurer l'existence d'une industrie laitière saine et viable au Canada.

3. Income Tax text (used by Tester 2 for scaled-up testing in Chapter 4)
Source text: from Canada Customs and Revenue Agency site

< http://www.ccra-adrc.gc.ca/tax/individuals/topics/income-tax/filing-obligations/menu-f.html >
accessed March 13, 2004

375 words

Devez-vous produire une déclaration de revenus?

Vous devez produire une déclaration de revenus pour 2003 si vous êtes dans **l'une** des situations suivantes :

vous avez de l'impôt à payer en 2003;

nous vous avons demandé de produire une déclaration;

vous avez disposé d'une immobilisation en 2003 (par exemple, vous avez vendu un terrain ou des actions) ou vous avez réalisé un gain en capital imposable (par exemple, un fonds commun de placement ou une fiducie vous a attribué un gain en capital, ou vous devez déclarer une provision pour gains en capital que vous avez demandée en 2002);

vous devez rembourser une partie ou la totalité des prestations de la Sécurité de la vieillesse ou des prestations d'assurance-emploi que vous avez reçues. Consultez la **ligne 235** – Remboursement des prestations de programmes sociaux;

vous n'avez pas remboursé la totalité des montants que vous avez retirés de votre régime enregistré d'épargne-retraite dans le cadre du Régime d'accession à la propriété ou du Régime d'encouragement à l'éducation permanente. Pour en savoir plus, consultez le guide intitulé **RC4135 – Régime d'accession à la propriété** ou le guide **RC4112 - Régime d'encouragement à l'éducation permanente (REEP)**;

vous devez cotiser au Régime de pensions du Canada parce que, en 2003, le total de vos revenus nets tirés d'un travail indépendant et de vos revenus d'emploi donnant droit à pension dépasse 3 500 $. Consultez la **ligne 222** – Déduction pour cotisations au RPC ou au RRQ pour le revenu d'un travail indépendant et pour d'autres revenus.

Remarque

Même si vous n'êtes pas dans l'une de ces situations, vous pouvez quand même produire une déclaration de revenus. Pour recevoir certaines prestations et crédits, vous devez produire une déclaration de revenus. Lisez **Avantages liés à la production d'une déclaration de revenus**.

Quoi de neuf

Nous vous verserons des intérêts composés quotidiennement sur votre remboursement d'impôt de 2003 à partir de la **plus rapprochée** des dates suivantes :

le 31 mai 2004;

le 31e jour après la date où vous avez envoyé votre déclaration;

le jour suivant la date où vous avez payé de l'impôt en trop.

Dates importantes

En général, vous devez nous envoyer votre déclaration de 2003 **au plus tard le 30 avril 2004**.

4. Health Care text (used by Tester 3 for scaled-up testing in Chapter 4)
Source text: from Health Canada site
< http://www.hc-sc.gc.ca/francais/soins/index.html >
accessed March 13, 2004

371 words

Soins de santé

Le financement du système de santé canadien relève en grande partie de l'État. Notre régime national d'assurance-maladie est une mosaïque de treize régimes provinciaux et territoriaux d'assurance-maladie, liés par le respect des principes nationaux établis au niveau fédéral.

La *Loi canadienne sur la santé* impose des critères et des conditions relatifs aux services de santé assurés et aux services de santé élargis que les provinces et les territoires doivent respecter pour recevoir la contribution financière complète au titre du Transfert canadien en matière de santé et de programmes sociaux (CHST). La *Loi canadienne sur la santé* vise à assurer à tous les citoyens admissibles du Canada un accès raisonnable aux soins de santé assurés, médicalement nécessaires et prépayés, sans avoir à payer directement au point de service.

Une responsabilité partagée

Le gouvernement fédéral, les dix provinces et les trois territoires ont des rôles primordiaux à jouer dans le système de soins de santé au Canada. Le gouvernement fédéral a les responsabilités suivantes :

établir et appliquer des principes ou des normes à l'échelle nationale pour le système de santé dans le cadre de la *Loi canadienne sur la santé* (pour de plus amples renseignements, veuillez consulter le site Web de la *Loi canadienne sur la santé*);

aider au financement des services provinciaux de soins de santé, par l'entremise de paiements de transferts fiscaux ;

offrir des services de santé directement à certains groupes tels que les anciens combattants, les Canadiens autochtones vivant sur les réserves, le personnel militaire, les détenus des pénitenciers fédéraux et la Gendarmerie royale du Canada ; et

d'assumer d'autres fonctions relatives à la santé comme la protection de la santé, la prévention des maladies et la promotion de la santé.

Les gouvernements provinciaux et territoriaux ont les responsabilités suivantes :

gérer et dispenser les services de santé assurés (pour de plus amples renseignements sur ces régimes d'assurance-maladie, veuillez consulter les textes soumis par les provinces et les territoires pour le rapport annuel sur la *Loi canadienne sur la santé*);

planifier, financer et évaluer les prestations de soins hospitaliers, des services médicaux et paramédicaux ; et de

gérer certains aspects des soins sur ordonnance et de santé publique.

5. Corporate Overview - National Research Council (used by Tester 4 for scaled-up testing in Chapter 4)
Source text: from National Research Council site
< http://www.nrc-cnrc.gc.ca/aboutUs/corporateoverview_f.html >
accessed March 18, 2004
385 words

Vue d'ensemble du CNRC

Structure

Le CNRC regroupe 19 instituts et programmes nationaux couvrant une multitude de domaines et une grande variété de services. Le CNRC a des bureaux dans toutes les provinces du Canada que et joue un rôle déterminant en alimentant l'innovation au sein des collectivités.

Les instituts et programmes du CNRC sont structurés en trois (3) grands secteurs :

le Programme de recherche en sciences physiques et du génie

le Programme de recherche en sciences de la vie et technologies de l'information

le Soutien technologique et industriel

Le Programme de recherche en sciences physiques et du génie comprend les quatre instituts de technologie de fabrication (CI-CNRC, ITFI-CNRC, ITPCE-CNRC et IMI-CNRC) et les quatre instituts (IRA-CNRC, IRC-CNRC, ITO-CNRC et IHA-CNRC).

Le Programme de recherche Sciences de la vie et technologies de l'information, portefeuille qui regroupe les cinq instituts de biotechnologie (IRB-CNRC, ISB-CNRC, IBP-CNRC, IBD-CNRC, IBM-CNRC), l'initiative en génomique et en santé, l'ISSM-CNRC, l'INN-CNRC, l'IENM-CNRC et les instituts du secteur des technologies de l'information et des communications, l'ISM-CNRC et l'ITI-CNRC.

Le Soutien technologique et industriel rassemble divers services destinés à l'industrie, notamment l'Institut canadien de l'Information scientifique et technique et le Programme d'aide à la recherche industrielle du CNRC, qui met l'accent sur les besoins des petites et moyennes entreprises.

Régie

Le CNRC est un organisme du gouvernement du Canada. Il rend des comptes au Parlement par le truchement du ministre de l'Industrie et est administré par un conseil de 22 membres choisis parmi sa clientèle. Pour connaître la composition du conseil d'administration, veuillez activer le lien ci-dessous ou cliquer l'onglet « Conseil d'administration » sur la barre de navigation, à gauche.

Pour en savoir plus :

Conseil d'administration
Liste des membres actuels du Conseil d'administration du CNRC.

Organigramme du CNRC
La structure organisationnelle du Conseil national de recherches (instituts, directions et programmes), ainsi que l'équipe de gestion.

Biographie des cadres
Apprenez-en davantage sur la haute direction du CNRC. Vous y trouverez une biographie succincte du président, des vice-présidents et de la secrétaire générale du CNRC, ainsi que la liste des directeurs généraux des instituts, programmes et directions générales du Conseil.

Capsules
Apprenez-en davantage sur le CNRC, notamment ses effectifs, son budget, l'emplacement de ses services au Canada, ses réalisations historiques, la *Loi sur le CNRC* et les archives du Conseil.

Appendix B Human Translations

1. SARS text

Human translation: text: taken from Health Canada Online
<http://www.hc-sc.gc.ca/english/protection/warnings/sars/questions6.html>
accessed April 7, 2003

283 words

Questions and Answers - Severe Acute Respiratory Syndrome (SARS)

Workplace health

Q. How is Health Canada supporting federal employees with respect to SARS?

A. Health Canada's Workplace Health and Public Safety Programme is providing advice to federal employers on occupational health and safety matters. Health professionals from Health Canada's Workplace Health and Public Safety Programme were on site at the Pearson, Vancouver and Dorval airports to provide information sessions for federal employees. They continue to provide information sessions to employers and employees at workplaces across the country, and answer daily enquiries from employers.

The Workplace Health and Safety Programme has issued advisories and has spoken to or met with representatives of a number of federal departments to provide them with information and advice they can use in determining what health and safety measures they deem appropriate for their employees. WHPSP officials participated in a videoconference with employees working in seven missions in affected areas overseas to provide information about respiratory protection and to discuss other general health concerns about SARS.

Q. Why is Health Canada not recommending the use of masks by federal workers, such as Customs agents at Pearson and Vancouver airports?

A. The role of Health Canada's Workplace Health and Public Safety Programme (WHPSP) is to provide advice to federal employers on occupational health and safety matters.

It is Health Canada's position that Government of Canada workers are unlikely to become ill because employees' encounters with passengers/clients are of reasonably short duration and not considered to be close contact. Therefore, Health Canada advises that the use of masks is not necessary.

Employers can use this information and advice in determining what health and safety measures they deem appropriate for their employees.

2. Canadian Dairy Commission text

from Canadian Dairy Commission site
< http://www.cdc.ca/cdc/main_e.asp?catid=663&page=1760 >
accessed March 13, 2004

278 words

History

Since the appointment of the
first Dominion Dairy Commissioner in 1890, the federal
government has played an active role in the development
and implementation of policies and programs in support of the
dairy industry.

Early Initiatives

Early federal initiatives included an iced butter railway car service (1895); funding for cool cheese curing rooms (1902); cow testing programs (1902); and the grading of butter and cheese for export (1923). The Government of Canada introduced a temporary subsidy on cheese and butter in 1935. Further programs to support prices, export surplus products, and limit imports were in effect throughout the 1940's and 1950's and were extended with the establishment of the Agricultural Stabilization Board in 1958.

Although it provided the necessary structure for price stabilization operations, the Agricultural Stabilization Board was not in a position to tackle two major problems critically affecting the dairy sector:

- a lack of co-ordination between federal and provincial policies, and
- the absence of an effective mechanism to control milk production.

It became clear that a mechanism was needed to ensure reasonable support prices for storable dairy products to underpin producer returns. New mechanisms were also needed to control costly surplus production. The federal government convened the 1963 Canadian Dairy Conference, which led to the creation of the Canadian Dairy Advisory Committee the same year. In its 1965 final report, this Committee recommended the creation of the Canadian Dairy Commission.

The Canadian Dairy Commission (CDC) was established with the proclamation of the *Canadian Dairy Commission Act* on October 31, 1966. Through the creation of this special Crown corporation, the federal government confirmed its commitment to a strong and viable dairy industry in Canada.

3. Income Tax text
from Customs Canada and Revenue Agency site
< http://www.ccra-adrc.gc.ca/tax/individuals/topics/income-tax/filing-obligations/menu-e.html >
accessed March 13, 2004

305 words

Do you have to file a tax return?

You have to file a tax return for 2003 if **any** of the following applies:

You have to pay tax for 2003.

We sent you a request to file a tax return.

You disposed of property in 2003 (for example, if you sold real estate or shares) or you realized a taxable capital gain (for example, if a mutual fund or trust attributed amounts to you, or you are reporting a capital gains reserve you claimed on your 2002 return).

You have to pay back any of your Old Age Security or Employment Insurance benefits. See **line 235** – Social benefits repayment, for details.

You have not repaid all of the amounts you withdrew from your registered retirement savings plan (RRSP) under the Home Buyers' Plan or the Lifelong Learning Plan. For details, see the **Home Buyers' Plan** or **Lifelong Learning Plan (LLP)** guides.

You have to contribute to the Canada Pension Plan (CPP). This can apply if, for 2003, the total of your net self-employment income and pensionable employment income is more than $3,500. See **line 222** – Deduction for CPP or QPP contributions on self-employment and other earnings.

Note

Even if none of these requirements applies, you may still want to file a tax return. To receive certain benefits and credits, you have to file a tax return. See **Benefits of filing a tax return**.

What's new

We will pay you compound daily interest on your tax refund for 2003 starting on whichever of the following three dates is **latest**:

May 31, 2004;

the 31st day after you file your tax return; or

the day after you overpaid your taxes.

Important dates

Generally, your tax return for 2003 has to be filed **on or before April 30, 2004**.

4. Health Care text
from Health Canada site
< http://www.hc-sc.gc.ca/english/care/index.html >
accessed March 13, 2004

295 words

Health Care

Canada has a predominantly publicly financed health care system. Our national health insurance program is achieved through thirteen interlocking provincial and territorial health insurance plans, linked through adherence to national principles set at the federal level.

The Canada Health Act establishes criteria and conditions related to insured health care services and extended health care services that the provinces and territories must meet in order to receive the full federal cash contribution under the Canada Health and Social Transfer (CHST). The aim of the Canada Health Act is to ensure that all eligible residents of Canada have reasonable access to medically necessary insured services on a prepaid basis, without direct charges at the point of service.

A shared responsibility

The federal government, the ten provinces, and the three territories have key roles to play in the health care system in Canada.

The federal government is responsible for:

setting and administering national principles or standards for the health care system through the Canada Health Act (for more information, visit the **Canada Health Act Web site**);

assisting in the financing of provincial health care services through fiscal transfers;

delivering direct health services to specific groups including veterans, native Canadians, persons living on reserves, military personnel, inmates of federal penitentiaries and the Royal Canadian Mounted Police;

fulfilling other health-related functions such as health protection, disease prevention, and health promotion.

The provincial and territorial governments are responsible for:

managing and delivering insured health services (for more information on these health care insurance plans, visit the provincial and territorial submissions to the **Canada Health Act Annual Report**);

planning, financing, and evaluating the provision of hospital care, physician and allied health care services, and

managing some aspects of prescription care and public health.

5. Corporate Overview – National Research Council Canada
from National Research Council site
< http://www.nrc-cnrc.gc.ca/aboutUs/corporateoverview_e.html>
accessed March 18, 2004
319 words

Corporate Overview

Organization

NRC is composed of 19 different institutes and national programs, spanning a wide variety of disciplines and offering a broad array of services. We are located in every province in Canada and play a major role in stimulating community-based innovation.

NRC institutes and programs are organized into three (3) key areas:

Physical Sciences and Engineering

Life Sciences and Information Technology

Technology and Industry Support

The Physical Sciences and Engineering Research Program includes the four manufacturing technology institutes (NRC-IC, NRC-IMTI, NRC-ICPET and NRC-IMI), and the following four institutes (NRC-IAR, NRC-IRC, NRC-IOT and NRC-HIA).

The Life Sciences and Information Technology is a portfolio comprised of the five biotechnology institutes (NRC-BRI, NRC-IBS, NRC-PBI, NRC-IBD, NRC-IMB) and the Genomics and Health Initiative, NRC-SIMS, NRC-NINT, NRC-INMS and the information and communications technology institutes of NRC-IMS and NRC-IIT.

Technology and Industry Support includes a number of industry-facing services, such as the NRC Canada Institute for Scientific and Technical Information and the NRC Industrial Research Assistance Program, which focuses on the needs of small- and medium-sized businesses.

Governance

NRC is an agency of the Government of Canada, reporting to Parliament through the Minister of Industry. It is governed by a council of 22 appointees drawn from its client community. For a list of current Council members, click on the link below, or choose the Council heading from the left-hand navigation bar.

For Further Information

NRC Council Members
Provides a listing of current members of the NRC Governing Council.

NRC Organizational Chart
NRC's organizational chart (institutes, branches and programs) as well as the NRC Executive Team.

Executive Bios
Find out more about NRC's executive team. Includes short biographies of the NRC President, Vice-Presidents, and Secretary-General and a listing of Directors-General of each NRC Institute, Program and Branch.

NRC Fast Facts
Find out more about the NRC, including staff size, budget, locations across Canada, historical achievements, the NRC Act and NRC Archives.

Appendix C Raw Output

1. SARS text - Systran

5 paragraphs (1heading, 2 questions and 2 responses)
12 sentences
361 words

Questions and answers - severe acute respiratory Syndrome (SRAS)
Health at the work

Q How Santé Canada does it support the federal civils servant in what milked with the SRAS?

R. The role of the Programme of health at the work and safety of public (PSTSP) of Canada Health consists in advising the federal civils servant in connection with the questions of professional hygiene and safety. The professionals of the health of the PSTSP of Canada Health were on the spot with the Pearson airports, of Vancouver and Dorval to organize meetings of information for the federal civils servant. They continue to organize meetings of information for the employers and the employees in work environment everywhere to the country, and to answer at the daily requests of information on behalf of the employers.

The Programme of health at the work and safety of the public gave opinions, and members of its personnel discussed and joined together with representatives a certain number of federal ministries in order to provide them information and councils to the measures of health and safety considered adapted for their employees. The representatives of the PSTSP attended a videoconference in the presence of the employees working in seven missions located abroad in the areas touched in order to provide them information on the protection of the respiratory tracts and to discuss other concerns of general health about SRAS.

Q Why Santé Canada doesn't it recommend the use of masks by the federal civils servant such as the agents of customs to the Pearson airports and Vancouver?

R. The role of the PSTSP of Canada Health consists in providing councils to the federal civils servant on questions of professional hygiene and safety.

Santé Canada believes that the employees of the federal government are not likely to become sick, for the contacts of the employees with the passengers or the customers are one relatively short duration and are not regarded as narrow. Consequently, Santé Canada considers that the use of the masks is not necessary.

The employers can use this information and these councils in order to determine measurements of health and safety which they consider suitable for their employees.

1. SARS text - ReversoPro
5 paragraphs (1 heading, 2 questions and 2 responses)
12 sentences
360 words

Questions and answers - austere pointed respiratory Syndrome (SRAS)
health to the work
Q. How Health Canada does it{*he*} rest{*does support*} the federal civil servants in what concerns the SRAS?
R. The role of the Program of health in the work and safety{*security*} of the public (PSTSP) of Health Canada consists in advising the federal civil servants about the questions of professional hygiene and safety{*security*}. The professionals of the health of the PSTSP of Health Canada were on the spot in airports Pearson, of Vancouver and Dorval to organize sessions of information in aid of the federal civil servants. They continue to organize sessions of information for the employers and the employees in working circles everywhere in the country, and to answer the daily demands of information on behalf of the employers. The Program of health in the work and safety{*security*} of the public emitted{*uttered*} opinions, and members of his{*her*} staff spoke and met with representatives of a certain number of federal ministries to supply them with the information and with advices about the measures of health and safety{*security*} considered suited for their employees. The representatives of the PSTSP attended a videoconference in the presence of the employees working in seven missions abroad situated in regions got{*touched*} to supply them with the information about the protection of respiratory tracts and to discuss the other preoccupations of general health about SRAS.

Q. Why Health Canada does not he recommend the custom{*usage*} of masks by the federal civil servants such as the customs officers in airports Pearson and of Vancouver?

R. The role of the PSTSP of Health Canada consists in supplying advices to the federal civil servants on questions of professional hygiene and safety{*security*}. Health Canada believes that the employees of the federal government do not risk to become patients, because the contacts of the employees with the passengers or the customers are of relatively short duration and are not considered as narrow. Consequently, Health Canada considers that the custom{*usage*} of masks is not necessary. The employers can use this information and these advices to determine the measures of health and safety{*security*} which they consider suited for their employees.

2. Canadian Dairy Commission - MT System 1

341 words
15 sentences
5 paragraphs

History

Since the appointment of the first Police chief of the milk of the Dominion in 1890, the federal government played an active part in the development and the application of the policies and programs in order to support the milk industry.

First programs

The first programs included/understood a service of coaches cooled for butter in 1895, of the fresh rooms for the maturation of cheese in 1902, the tests for the milch cows in 1902 and the classification of butter and cheese intended for export in 1923. Later, in 1935, the government of Canada adopted a provisional subsidy for cheese and

butter. During Forties and Fifties, it launched programmes of maintenance prices, of export of surplus products and limitation of the imports, programs which were continued after the creation of the Office of stabilization of the farm prices in 1958.

Although it offered the desired structure for the activities of stabilization of the prices, the Office of stabilization of the farm prices was not able to attack two major problems which affected the dairy sector seriously:

1) lack of coordination between the federal and provincial policies, and

2) the absence of an effective mechanism of control of the dairy production.

Such a mechanism proved to be necessary to ensure of the prices of support reasonable for the dairy products to store in order to reinforce the incomes of the producers. New mechanisms were also necessary to control the surpluses of expensive production. In 1963, the federal government convened a Canadian Conference of the milk industry, which led to the creation of the Consultative Committee of the milk industry, the same year. The final report/ratio of this Committee, presented in 1965, recommended the creation of the Canadian Commission of milk.

The Canadian Commission of milk (CCL) was established by the proclamation of the Law on the Canadian Commission of milk, October 31, 1966. By the creation of this company of the particular Crown, the federal government confirmed its commitment to ensure the existence of a healthy and viable milk industry Canada.

2. Canadian Dairy Commission text - MT System 2
333 words
15 sentences
5 paragraphs

Historic

Since the appointment of the first Police superintendent of the milk of the Dominion in 1890, the federal government played an active role in the settling and the application of politics{*policies*} and programs to rest{*support*} the dairy farming.

First programs

First programs understood{*included*} a service of cars cooled for the butter in 1895, fresh{*cool*} rooms for the maturation of the cheese in 1902, tests for dairy cows in 1902 and classification of the butter and the cheese intended for the export in 1923. Later, in 1935 , the government of Canada adopted a temporary subsidy for the cheese and the butter. During the forties and fifty, it{*he*} threw{*launched*} programs of support of prices, export of superfluous products and limitation of imports, schedule{*program*} which{*who*} were continued after the creation of the Office of price stabilization agricultural in 1958.

Although it{*he*} offered structure wanted for the activities of price stabilization, the Office of price stabilization agricultural was not capable of attacking{*affecting*} two major problems which affected seriously the sector milkman:

- 1) the lack of coordination between federal and provincial politics{*policies*}, and
- 2) the absence of an effective mechanism of control of the production dairywoman.

Such a mechanism turned out necessary to assure{*insure*} reasonable support prices for dairy products to be stored to strengthen the incomes of the producers. New mechanisms were also required to control the expensive surpluses of production. In 1963, the federal government summoned{*convened*} a Canadian Conference of the dairy farming, what led to the creation of the consultative Committee of the dairy farming, the same year. The final report of this Committee, presented in 1965, recommended the creation of the Canadian Committee{*commission*} of the milk.

The Canadian Committee{*commission*} of the milk (CCL) was established by the proclamation of the Law on the Canadian Committee{*commission*} of the milk, on October 31, 1966. By the creation of this company{*society*} of the particular Crown, the federal government confirmed its commitment to assure{*insure*} the existence of a healthy and viable dairy farming in Canada.

3. Income Tax text - MT System 1

368 words
22 sentences
4 paragraphs

Do you have to produce an income tax return?

You must produce an income tax return for 2003 if you are in one of the following situations:

· you have tax to pay into 2003;

· we asked you to produce a declaration;

· you had an immobilization into 2003 (for example, you sold a ground or actions) or you carried out a taxable capital gain (for example, investment funds or a trust allotted a capital gain to you, or you must declare a provision for capital gains which you asked into 2002);

· you must refund a part or totality of the services of the Safety of old age or services of insurance-employment which you received. Consult line 235 - Refunding of the services of social programs;

· you did not refund the totality of the amounts which you withdrew from your recorded mode of saving-retirement within the framework of the Mode of possibility of home-ownership or the Mode of encouragement to further education. Of to know more, consult the guide entitled RC4135 - Mode of possibility of home-ownership or guide RC4112 - Mode of encouragement to further education (REEP);

· you owe cotiser with the pension Scheme of Canada because, in 2003, the total of your clear incomes drawn from a work independent and of

your incomes of employment giving right pension exceeds 3 500 $. Consultez line 222 - Deduction for contributions with the RPC or the RRQ for the income of an independent work and for other incomes.

Notice

Even if you are not in one of these situations, you can nevertheless produce an income tax return. To receive certain services and appropriations, you must produce an income tax return. Read Avantages related to the production of an income tax return.

What is nine

We will pour you interests made up daily on your refunding of tax of 2003 from close to the following dates:

· on May 31 2004;

· the 31e day after the date where you sent your declaration;

· the day following the date where you paid tax in excess.

Important dates

In general, you must send your declaration of 2003 to us at the latest on April 30 2004.

3. Income Tax text - MT System 2
392 words
22 sentences
4 paragraphs

Do you have to produce a tax return?
You have to produce a tax return for 2003 if you are in one of the following situations:

- you have of the tax to be paid in 2003;
- we asked you to produce a statement;
- you had an immobilization in 2003 (for example, you sold a ground or actions) or you realized an earning in taxable capital (for example, a mutual fund or a trust attributed{*awarded*} you an earning in capital, or you have to declare a reserve for earnings in capital for which you asked in 2002);
- you have to pay off a part or the totality of the performances of the Safety{*security*} of the old age or the performance of assurance - use which you received. Consult the line 235 - Refund{*repayment*} of the performances of social programs;
- you did not pay off the totality of the amounts which you removed from your regime registered{*recorded*} with retirement savings scheme within the framework of the Regime of property ownership or the Regime of encouragement in continuing education. To know more about it, consult the entitled guide-'RC4135 RC4135 regime daccession in the property or the guide " RC4112 - Regime dencouragement in léducation permanent (REEP);
- you have to pay the contribution in the Regime of pensions of Canada because, in 2003 , the total of your net incomes pulled{*fired*} by an independent work and by your incomes of pensionable use{*job*} exceeds 3 500 $. Consult the line 222 - Deduction for contributions in the RPC or in the RRQ for the income of an independent work and for the other incomes.

Remark
Even though you are not in one of these situations, you can all the same produce a tax return. To receive certain performances and credits, you have to produce a tax return. Read ' Advantages bound{*connected*} to the production dune tax return.

What's new
We shall pay{*pour*} you compound interests daily on your refund{*repayment*} of tax of 2003 from the most moved closer of the next dates:

- on May 31, 2004;
- the 31-th day after date when you sent your statement;
- the day following the date when you paid of the tax too.

Important dates
Generally, you have to send us your statement of 2003 at the latest on April 30, 2004.

4. Health Care text - MT System 1
359 words
16 sentences
4 paragraphs

Care of health

The financing of the Canadian system of health concerns the State mainly. Our national scheme of health insurance is a mosaic of thirteen provincial and territorial schemes of health insurance, bound by the respect of the national principles established at the federal level.

The Canadian Law on health imposes criteria and conditions relating to the assured health services and the widened health services that the provinces and the territories must respect to receive the complete financial contribution to the title of the Canadian Transfer health social programs (CHST). The Canadian Law on health aims at ensuring all the acceptable citizens of Canada a reasonable access to the ensured care of health, médicalement necessary and prepaid, without having to pay directly at the point service.

A shared responsibility

The federal government, the ten provinces and the three territories have roles paramount to play in the system of care of health in Canada. The federal government has the following responsibilities:

· to establish and apply principles or standards on a national scale for the system of health within the framework of the Canadian Law to health (for further information, please consult the Web site of the Canadian Law on health);

· to contribute to the financing of the provincial services of care of health, via payments of tax transfers;

· to offer health services directly to certain groups such as ex-serviceman, the Canadians autochtones alive on the reserves, the military personnel, the prisoners of the federal penitentiaries and the royal Gendarmerie of Canada; and

· to take up other duties relating to health like the protection of health, the prevention of the diseases and the promotion of health.

The provincial and territorial governments have the following responsibilities:

· to manage and exempt the assured health services (for further information on these schemes of health insurance, please consult the texts subjected by the provinces and the territories for the annual report on the Canadian Law on health);

· to plan, finance and evaluate the services of hospital care, the medical departments and ancillary medical; and of

· to manage certain aspects of the care on ordinance and public health.

4. Health Care text - MT System 2
368 words

16 sentences
4 paragraphs

Care of health

The financing of the Canadian system of health recovers in big part of the State. Our national regime of health insurance is a mosaic of thirteen provincial and territorial regimes of health insurance, bound{*connected*} by the respect for national principles established at federal level.

Canadian Law on the health imposes criteria and conditions relatives on the assured{*insured*} health services and on the widened health services which provinces and territories have to respect to receive complete financial contribution in conformance with the Canadian Transfer in health and in social programs (CHST). Canadian Law on the health aims to assure{*insure*} to all the acceptable citizens of Canada a reasonable access to the assured{*insured*} care of health, medically necessary and prépayés, without having directly to pay to the point of service.

A shared responsibility
The federal government, ten provinces and three territories have roles original to play in the system of care of health in Canada. The federal government has the following responsibilities:

- establish and apply principles or standards on a national scale for the system of health within the framework of the Canadian Law to the health (should you require further information, please consult the Web site of the Canadian Law on the health;
- to help in the financing of the provincial services of care of health, by the intervention of payments of fiscal transfers;
- to offer health services directly to certain groups such as the war veterans, the autochtonous Canadians living on reserves, military staff, prisoners of the federal penitentiaries and the royal Police force of Canada; and
- to assume the other functions{*offices*} relative to the health as the protection of the health, the prevention of diseases and the promotion of the health.

The provincial and territorial governments have the following responsibilities:

- administer and distribute assured{*insured*} health services (should you require further information on these regimes of health insurance, please consult texts subjected with provinces and territories for the annual report on the Canadian Law on the health;
- to plan, fork out and to estimate the performances of hospitable care, medical and ancillary medical services; and of
- to administer certain aspects of care on prescription and health service.

5. National Research Council - Corporate Overview text - MT System 1
400 words
25 sentences
10 paragraphs

Overall picture of the CNRC

Structure

The CNRC gathers 19 national institutes and programs covering a multitude of fields and a large variety of services. The CNRC has offices in all the provinces of Canada that and plays a part determining by feeding the innovation within the communities.

The institutes and programs of the CNRC are structured in three (3) great sectors:

§ the genius and research Program in physical sciences
§ the research Programme in life sciences and information technologies
§ the technological and industrial Support

The genius and research Program in physical sciences includes/understands the four institutes of technology of manufacture (CI-CNRC, ITFI-CNRC, ITPCE-CNRC and IMI-CNRC) and four institutes (IRA-CNRC, IRC-CNRC, ITO-CNRC and IHA-CNRC).

The research Program Life sciences and information technologies, wallet which gathers the five institutes of biotechnology (IRB-CNRC, ISB-CNRC, IBP-CNRC, IBD-CNRC, IBM-CNRC), the initiative into genomic and health, the ISSM-CNRC, the lNN-NRC, the IENM-CNRC and institutes of the sector of communication and information technologies, the ISM-CNRC and the ITI-CNRC.

The technological and industrial Support gathers various services intended for industry, in particular the Canadian Institute of scientific and technical Information and the Supplementary programme to the industrial research of the CNRC, which stresses the needs for small and medium-sized undertakings.

Control

The CNRC is an organization of the government of Canada. It returns accounts to the Parliament by the means of the Minister of Industry and is managed by a council of 22 members chosen among his customers.

To know the composition of the board of directors, please activate the bond below or cliquer the mitre " Board of directors " on the bar of navigation, on the left.

Of to know more:

Board of directors
List current members of the Board of directors of the CNRC.

Flow chart of the CNRC
The organisational structure of the national Council of research (institutes, directions and programs), as well as the team of management.

Biography of the frameworks
Learn more on the top management from the CNRC. You will find there a biography brief of the president, vice-presidents and of the secretary-general of the CNRC, as well as the list of the general managers of the institutes, programs and general directions of the Council.

Capsules
Learn more on the CNRC, in particular its manpower, its budget, the site of its services in Canada, its achievements historical, the Law on the CNRC and the files of the Council.

5. National Research Council – Corporate Overview text - MT System 2
403 words
25 sentences
10 paragraphs

General view of the CNRC

Structure

The CNRC regroups 19 institutes and national programs covering a multitude of domains and a big variety of services. The CNRC has desks in all the provinces of Canada which and plays a determining role by feeding innovation within communities.

Institutes and programs of the CNRC are structured in three (3) big sectors:

- § the Program of search in physical sciences and the genius
- § the Program of search in life sciences and information technologies
- § technological and industrial Support

The Program of search in physical sciences and the genius understands{*includes*} the four institutes of technology of manufacture (CI-CNRC, ITFI-CNRC, ITPCE-CNRC AND IMI-CNRC) and four institutes (IRA-CNRC, IRC-CNRC, ITO-CNRC AND IHA-CNRC).

The Program of search Life sciences and information technologies, wallet which regroups the five institutes of biotechnology (IRB-CNRC, ISB-CNRC, IBP-CNRC, IBD-CNRC, IBM-CNRC), the initiative in génomique and in health, the ISSM-CNRC, the INN-CNRC, the IENM-CNRC and the institutes of the sector of information technologies and communications, ISM-CNRC and ITI-CNRC.

Technological and industrial Support collects different services intended for the industry, notably the Canadian Institute of the scientific and technical Information and the Program of help{*assistant*} in the industrial search for the CNRC, which emphasizes the necessities of small and medium-sized companies.

State control

The CNRC is a body of the government of Canada. It{*he*} returns accounts to the Parliament with the aid of Secretary of Commerce and is administered by an advice{*council*} of 22 members chosen among the clientele. To know the composition of the board of directors, please activate link below or click the thumb index "Board of directors" the bar of navigation, to the left.

To know more about it:

' Council dadministration
List current members of the Board of directors of the CNRC.

Organization chart of the CNRC
The organizational structure of the national council of searches (institutes, directions{*managements*} and programs), as well as the team of management.

Biography of the executives
Learn there more on the high direction{*management*} of the CNRC. You will find there a brief biography of the president, the vice-presidents and the general secretary of the CNRC, as well as the list of the general managers of institutes, programs and head offices of the council.

Capsules
Learn there more on the CNRC, notably the actual, its budget, the place of its services in Canada, its historic realizations, Law on the CNRC and the archives of the council.

Appendix D Scored Raw Output

1. SARS text - Systran

5 paragraphs (1 heading, 2 questions and 2 responses)
12 sentences
361 words

[NB errors are underlined]

Questions and answers - severe acute respiratory Syndrome (SRAS)
Health at the work

Accuracy: 2
Intelligibility: 2

Coherence for this paragraph: 2

Q How Santé Canada does it support the federal civils servant in what milked with the SRAS?

Accuracy: 4
Intelligibility: 5

Coherence for this paragraph: 3

R. The role of the Programme of health at the work and safety of public (PSTSP) of Canada Health consists in advising the federal civils servant in connection with the questions of professional hygiene and safety.

Accuracy: 3
Intelligibility: 3

The professionals of the health of the PSTSP of Canada Health were on the spot with the Pearson airports, of Vancouver and Dorval to organize meetings of information for the federal civils servant.

Accuracy: 3
Intelligibility: 3

They continue to organize meetings of information for the employers and the employees in work environment everywhere to the country, and to answer at the daily requests of information on behalf of the employers.

Accuracy: 2
Intelligibility: 3

The Programme of health at the work and safety of the public gave opinions, and members of its personnel discussed and joined together with representatives a certain number of federal ministries in order to provide them information and councils to the measures of health and safety considered adapted for their employees.

Accuracy: 3

Intelligibility: 3

The representatives of the PSTSP attended a videoconference in the presence of the employees working in seven missions located abroad in the areas touched in order to provide them information on the protection of the respiratory tracts and to discuss other concerns of general health about SRAS.

Accuracy: 3
Intelligibility: 2

Coherence for this paragraph: 2

Q Why Santé Canada doesn't it recommend the use of masks by the federal civils servant such as the agents of customs to the Pearson airports and Vancouver?

Accuracy: 2
Intelligibility: 3

Coherence for this paragraph: 2

R. The role of the PSTSP of Canada Health consists in providing councils to the federal civils servant on questions of professional hygiene and safety.

Accuracy: 3
Intelligibility: 3

Santé Canada believes that the employees of the federal government are not likely to become sick, for the contacts of the employees with the passengers or the customers are one relatively short duration and are not regarded as narrow.

Accuracy: 3
Intelligibility: 3

Consequently, Santé Canada considers that the use of the masks is not necessary.

Accuracy: 2
Intelligibility: 2

The employers can use this information and these councils in order to determine measurements of health and safety which they consider suitable for their employees.

Accuracy: 3
Intelligibility: 2

Coherence for this paragraph: 2

Coherence for text as a whole: 2

Summary of scoring:

Accuracy: 2+4+3+3+2+3+3+2+3+3+2+3 = 33/12 sentences = 2.50

Intelligibility: 2+5+3+3+3+3+2+3+3+3+2+2 = 34/12 sentences = 2.83

Coherence: 2+3+2+2+2+2 = 13/6 paragraphs (5 +1 for entire document) = 2.17

1. SARS text - ReversoPro
5 paragraphs (1 heading, 2 questions and 2 responses)
12 sentences
360 words

[NB errors are underlined]
Questions and answers - austere pointed respiratory Syndrome (SRAS)

health to the work

Accuracy: 2
Intelligibility: 2

Coherence for this paragraph: 2

Q. How Health Canada does it{*he*} rest{*does support*} the federal civil servants in what concerns the SRAS?

Accuracy: 2
Intelligibility: 3

Coherence for this paragraph: 2

R. The role of the Program of health in the work and safety{*security*} of the public (PSTSP) of Health Canada consists in advising the federal civil servants about the questions of professional hygiene and safety{*security*}.

Accuracy: 3
Intelligibility: 3

The professionals of the health of the PSTSP of Health Canada were on the spot in airports Pearson, of Vancouver and Dorval to organize sessions of information in aid of the federal civil servants.

Accuracy: 2
Intelligibility: 3

They continue to organize sessions of information for the employers and the employees in working circles everywhere in the country, and to answer the daily demands of information on behalf of the employers.

Accuracy: 2
Intelligibility: 3

The Program of health in the work and safety{*security*} of the public emitted{*uttered*} opinions, and members of his{*her*} staff spoke and met with representatives of a certain number of federal ministries to supply them with the information and with advices about the measures of health and safety{*security*} considered suited for their employees.

Accuracy: 3
Intelligibility: 3

The representatives of the PSTSP attended a videoconference in the presence of the employees working in seven missions abroad situated in regions got{*touched*} to supply them with the information about the protection of respiratory tracts and to discuss the other preoccupations of general health about SRAS.

Accuracy: 2
Intelligibility: 4

Coherence for this paragraph: 1

Q. Why Health Canada does not he recommend the custom{*usage*} of masks by the federal civil servants such as the customs officers in airports Pearson and of Vancouver?

Accuracy: 2
Intelligibility: 3

Coherence for this paragraph: 2

R. The role of the PSTSP of Health Canada consists in supplying advices to the federal civil servants on questions of professional hygiene and safety{*security*}.

Accuracy: 2
Intelligibility: 2

Health Canada believes that the employees of the federal government do not risk to become patients, because the contacts of the employees with the passengers or the customers are of relatively short duration and are not considered as narrow.

Accuracy: 2

Intelligibility: 4

Consequently, Health Canada considers that the custom{*usage*} of masks is not necessary.

Accuracy: 2
Intelligibility: 1

The employers can use this information and these advices to determine the measures of health and safety{*security*} which they consider suited for their employees.

Accuracy: 2
Intelligibility: 3

Coherence for this paragraph: 2

Coherence for text as a whole: 2

Summary of scoring:

Accuracy: 2+2+3+2+2+3+2+2+2+2+2+2 = 26 /12 sentences = 2.17

Intelligibility: 2+3+3+3+3+3+4+3+2+4+1+3 = 34/12 sentences = 2.83

Coherence: 2+2+1+2+2+2 = 11/6 paragraphs (5 +1 for entire document) =1.83

Tester 1 - Canadian Dairy Commission text - MT System 1

341 words
15 sentences
5 paragraphs

History

Intelligibility: 1/5
Accuracy: 1/4

Since the appointment of the first Police chief of the milk of the Dominion in 1890, the federal government played an active part in the development and the application of the policies and programs in order to support the milk industry.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 2/4

First programs

Intelligibility: 1/5
Accuracy: 1/4

The first programs included/understood a service of coaches cooled for butter in 1895, of the fresh rooms for the maturation of cheese in 1902, the tests for the milch cows in 1902 and the classification of butter and cheese intended for export in 1923.

Intelligibility: 4/5
Accuracy: 3/4

Later, in 1935, the government of Canada adopted a provisional subsidy for cheese and butter.

Intelligibility: 2/5
Accuracy: 2/4

During Forties and Fifties, it launched programmes of maintenance prices, of export of surplus products and limitation of the imports, programs which were continued after the creation of the Office of stabilization of the farm prices in 1958.

Intelligibility: 4/5
Coherence: 3/5
Accuracy: 3/4

Although it offered the desired structure for the activities of stabilization of the prices, the Office of stabilization of the farm prices was not able to attack two major problems which affected the dairy sector seriously:

Intelligibility: 3/5
Accuracy: 3/4

1) lack of coordination between the federal and provincial policies, and

Intelligibility: 2/5
Accuracy: 1/4

2) the absence of an effective mechanism of control of the dairy production.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 1/4

Such a mechanism proved to be necessary to ensure of the prices of support reasonable for the dairy products to store in order to reinforce the incomes of the producers.

Intelligibility: 3/5
Accuracy: 3/4

New mechanisms were also necessary to control the surpluses of expensive production.

Intelligibility: 2/5
Accuracy: 2/4

In 1963, the federal government convened a Canadian Conference of the milk industry, which led to the creation of the Consultative Committee of the milk industry, the same year.

Intelligibility: 3/5
Accuracy: 3/4

The final report/ratio of this Committee, presented in 1965, recommended the creation of the Canadian Commission of milk.

Intelligibility: 2/5
Coherence: 3/5
Accuracy: 2/4

The Canadian Commission of milk (CCL) was established by the proclamation of the Law on the Canadian Commission of milk, October 31, 1966.

Intelligibility: 3/5
Accuracy: 2/4

By the creation of this company of the particular Crown, the federal government confirmed its commitment to ensure the existence of a healthy and viable milk industry Canada.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 2/4

Coherence of text: 3/5

Summary of scoring
Intelligibility:
Total score:
= 1+2+1+4+2+4+3+2+2+3+2+3+2+3+3 = 37
Average score:
= 37 total score / 15 sentences = 2.47
Coherence:
Total score:
= 2+3+2+3+2+3 = 15
Average score:
= 15 total score / 6 (5 paragraphs + 1 for text) = 2.5
Accuracy:
Total score:
= 1+2+1+3+2+3+3+1+1+3+2+3+2+2+2 = 31
Average score:
= total score 31 / 15 sentences = 2.07

Tester 1 – Canadian Dairy Commission text - MT System 2
333 words
15 sentences
5 paragraphs

Historic

Intelligibility: 2/5
Accuracy: 1/4

Since the appointment of the first Police superintendent of the milk of the Dominion in 1890, the federal government played an active role in the settling and the application of politics{*policies*} and programs to rest{*support*} the dairy farming.

Intelligibility: 4/5
Coherence: 4 /5
Accuracy: 3/4

First programs

Intelligibility: 1/5
Accuracy: 1/4

First programs understood{*included*} a service of cars cooled for the butter in 1895, fresh{*cool*} rooms for the maturation of the cheese in 1902, tests for dairy cows in 1902 and classification of the butter and the cheese intended for the export in 1923.

Intelligibility: 4/5
Accuracy: 3/4

Later, in 1935 , the government of Canada adopted a temporary subsidy for the cheese and the butter.

Intelligibility: 2/5
Accuracy: 2/4

During the forties and fifty, it{*he*} threw{*launched*} programs of support of prices, export of superfluous products and limitation of imports, schedule{*program*} which{*who*} were continued after the creation of the Office of price stabilization agricultural in 1958.

Intelligibility: 5/5
Coherence: 4/5
Accuracy: 3/4

Although it{*he*} offered structure wanted for the activities of price stabilization, the Office of price stabilization agricultural was not capable of attacking{*affecting*} two major problems which affected seriously the sector milkman:

Intelligibility: 5/5
Accuracy: 4/4

1) the lack of coordination between federal and provincial politics{*policies*}, and

Intelligibility: 2/5

Accuracy: 2 /4

2) the absence of an effective mechanism of control of the production dairywoman.

Intelligibility: 2/5
Coherence: 5/5
Accuracy: 2/4

Such a mechanism turned out necessary to assure{*insure*} reasonable support prices for dairy products to be stored to strengthen the incomes of the producers.

Intelligibility: 3/5
Accuracy: 3/4

New mechanisms were also required to control the expensive surpluses of production.

Intelligibility: 2/5
Accuracy: 2/4

In 1963, the federal government summoned{*convened*} a Canadian Conference of the dairy farming, what led to the creation of the consultative Committee of the dairy farming, the same year.

Intelligibility: 3/5
Accuracy: 3/4

The final report of this Committee, presented in 1965, recommended the creation of the Canadian Committee{*commission*} of the milk.

Intelligibility: 3/5
Coherence: 3/5
Accuracy: 2/4

The Canadian Committee{*commission*} of the milk (CCL) was established by the proclamation of the Law on the Canadian Committee{*commission*} of the milk, on October 31, 1966.

Intelligibility: 4/5
Accuracy: 2/4

By the creation of this company{*society*} of the particular Crown, the federal government confirmed its commitment to assure{*insure*} the existence of a healthy and viable dairy farming in Canada.

Intelligibility: 3/5
Coherence: 3/5
Accuracy: 2/4

Coherence for text: 4/5

Summary of scoring:
Intelligibility:
Total score:

= 2+4+1+4+2+5+5+2+2+3+2+3+3+4+3 = 45
Average score:
= 45 total score / 15 sentences = 3.00

Coherence:

Total score:
= 4+4+5+3+3+4 = 23
Average score:
= 23 total score / 6 (5 paragraphs + text) = 3.83

Accuracy:

Total score:
= 1+3+1+3+2+3+4+2+2+3+2+3+2+2+2 = 35
Average score:
= 35 total score / 15 number of sentences = 2.33

Tester 2 - Income Tax text - MT System 1
368 words
22 sentences
4 paragraphs

Do you have to produce an income tax return?

Intelligibility: 1 /5
Accuracy: 2 /4

You must produce an income tax return for 2003 if you are in one of the following situations:

Intelligibility: 2 /5
Accuracy: 2 /4

· you have tax to pay into 2003;

Intelligibility: 2 /5
Accuracy: 2 /4

· we asked you to produce a declaration;

Intelligibility: 2 /5
Accuracy: 2 /4

· you had an immobilization into 2003 (for example, you sold a ground or actions) or you carried out a taxable capital gain (for example, investment funds or a trust allotted a capital gain to you, or you must declare a provision for capital gains which you asked into 2002);

Intelligibility: 5/5
Accuracy: 4 /4

· you must refund a part or totality of the services of the Safety of old age or services of insurance-employment which you received.

Intelligibility: 3/5
Accuracy: 3/4

Consult line 235 - Refunding of the services of social programs;

Intelligibility: 3/5
Accuracy: 3 /4

· you did not refund the totality of the amounts which you withdrew from your recorded mode of saving-retirement within the framework of the Mode of possibility of home-ownership or the Mode of encouragement to further education.

Intelligibility: 4/5
Accuracy: 4 /4

Of to know more, consult the guide entitled RC4135 - Mode of possibility of home-ownership or guide RC4112 – Mode of encouragement to further education (REEP);

Intelligibility: 4 /5
Accuracy: 4 /4

· you owe cotiser with the pension Scheme of Canada because, in 2003, the total of your clear incomes drawn from a work independent and of your incomes of employment giving right pension exceeds 3 500 $.

Intelligibility: 5 /5
Coherence: 3 /5
Accuracy: 4 /4

Consultez line 222 - Deduction for contributions with the RPC or the RRQ for the income of an independent work and for other incomes.

Intelligibility: 4 /5
Accuracy: 4 /4

Notice

Intelligibility: 2 /5
Accuracy: 2 /4

Even if you are not in one of these situations, you can nevertheless produce an income tax return.

Intelligibility: 2 /5
Accuracy: 2 /4

To receive certain services and appropriations, you must produce an income tax return.

Intelligibility: 2 /5
Accuracy: 3 /4

Read Avantages related to the production of an income tax return.

Intelligibility: 2/5
Coherence: 2 /5
Accuracy: 3 /4

What is nine

Intelligibility: 5 /5
Accuracy: 4 /4

We will pour you interests made up daily on your refunding of tax of 2003 from close to the following dates:

Intelligibility: 5 /5
Accuracy: 4 /4

· on May 31 2004;

Intelligibility: 1/5
Accuracy: 1 /4

· the 31e day after the date where you sent your declaration;

Intelligibility: 2 /5
Accuracy: 2 /4

· the day following the date where you paid tax in excess.

Intelligibility: 2 /5
Coherence: 4 /5
Accuracy: 3 /4

Important dates

Intelligibility: 1 /5
Accuracy: 1 /4

In general, you must send your declaration of 2003 to us at the latest on April 30 2004.

Intelligibility: 2 /5
Coherence: 3 /5
Accuracy: 3 /4

Coherence for text: 3 /5

Summary of scoring:
Intelligibility:
Total score:
= 1+2+2+2+5+3+3+4+4+5+4+2+2+2+2+5+5+1+2+2+1+2 = 61
Average score:
= 61 total score / 22 sentences = 2.77
Coherence:
Total score:
= 3+2+4+3+3 = 15
Average score:
= 15 total score / 5 (4 paragraphs + text) = 3.00
Accuracy:
Total score:
= 2+2+2+2+4+3+3+4+4+4+4+2+2+3+3+4+4+1+2+3+1+3 = 62
Average score:
= 62 total score / 22 sentences = 2.82

Tester 2 - Income Tax text - MT System 2
392 words
22 sentences
4 paragraphs

Do you have to produce a tax return?

Intelligibility: 1 /5
Accuracy: 2 /4

You have to produce a tax return for 2003 if you are in one of the following situations:

Intelligibility: 1 /5
Accuracy: 2 /4

- you have of the tax to be paid in 2003;

Intelligibility: 3 /5
Accuracy: 2 /4

- we asked you to produce a statement;

Intelligibility: 1 /5
Accuracy: 2 /4

- you had an immobilization in 2003 (for example, you sold a ground or actions) or you realized an earning in taxable capital (for example, a mutual fund or a trust attributed{*awarded*} you an earning in capital, or you have to declare a reserve for earnings in capital for which you asked in 2002);

Intelligibility: 4/5
Accuracy: 4 /4

- you have to pay off a part or the totality of the performances of the Safety{*security*} of the old age or the performance of assurance - use which you received.

Intelligibility: 4/5
Accuracy: 4 /4

Consult the line 235 - Refund{*repayment*} of the performances of social programs;

Intelligibility: 3/5
Accuracy: 3 /4

- you did not pay off the totality of the amounts which you removed from your regime registered{*recorded*} with retirement savings scheme within the framework of the Regime of property ownership or the Regime of encouragement in continuing education.

Intelligibility: 5 /5
Accuracy: 4 /4

To know more about it, consult the entitled guide-'RC4135 RC4135 regime daccession in the property or the guide " RC4112 - Regime dencouragement in léducation permanent (REEP);

Intelligibility: 5/5
Accuracy: 4 /4

- you have to pay the contribution in the Regime of pensions of Canada because, in 2003 , the total of your net incomes pulled{*fired*} by an independent work and by your incomes of pensionable use{*job*} exceeds 3 500 $.

Intelligibility: 4 /5
Accuracy: 4 /4

Consult the line 222 - Deduction for contributions in the RPC or in the RRQ for the income of an independent work and for the other incomes.

Intelligibility: 3 /5
Coherence: 3 /5
Accuracy: 3 /4

Remark

Intelligibility: 2/5
Accuracy: 2 /4

Even though you are not in one of these situations, you can all the same produce a tax return.

Intelligibility: 1 /5
Accuracy: 2 /4

To receive certain performances and credits, you have to produce a tax return.

Intelligibility: 2 /5
Accuracy: 2 /4

Read ' Advantages bound{*connected*} to the production dune tax return.

Intelligibility: 2/5
Coherence: 3 /5
Accuracy: 3 /4

What's new

Intelligibility: 1 /5
Accuracy: 1 /4

We shall pay{*pour*} you compound interests daily on your refund{*repayment*} of tax of 2003 from the most moved closer of the next dates:

Intelligibility: 3 /5
Accuracy: 4 /4

- on May 31, 2004;

Intelligibility: 1/5
Accuracy: 1 /4

- the 31-th day after date when you sent your statement;

Intelligibility: 2 /5
Accuracy: 2 /4

- the day following the date when you paid of the tax too.

Intelligibility: 3 /5
Coherence: 2 /5
Accuracy: 3 /4

Important dates

Intelligibility: 1 /5
Accuracy: 1 /4

Generally, you have to send us your statement of 2003 at the latest on April 30, 2004.

Intelligibility: 3 /5
Coherence: 2 /5
Accuracy: 3 /4

Coherence for text: 3 /5

Summary of scoring:
Intelligibility:
Total score:
= 1+1+3+1+4+4+3+5+5+4+3+2+1+2+2+1+3+1+2+3+1+3 = 55
Average score:
= 55 total score / 22 sentences = 2.50
Coherence:
Total score:
= 3+3+2+2+3 = 13
Average score:
= 13 total score / 5 (4 paragraphs + text) = 2.60
Accuracy:
Total score:
= 2+2+2+2+4+4+3+4+4+4+3+2+2+2+3+1+4+1+2+3+1+3 =58
Average score:
= 58 total score / 22 sentences = 2.64

Tester 3 - Health Care text – MT System 1
359 words
16 sentences
4 paragraphs

Care of health

Intelligibility: 2/5
Accuracy: 1/4

The financing of the Canadian system of health concerns the State mainly.

Intelligibility: 2/5
Accuracy: 1/4

Our national scheme of health insurance is a mosaic of thirteen provincial and territorial schemes of health insurance, bound by the respect of the national principles established at the federal level.

Intelligibility: 2/5
Coherence: 1/5
Accuracy: 2/4

The Canadian Law on health imposes criteria and conditions relating to the assured health services and the widened health services that the provinces and the territories must respect to receive the complete financial contribution to the title of the Canadian Transfer health social programs (CHST).

Intelligibility: 3/5
Accuracy: 4/4

The Canadian Law on health aims at ensuring all the acceptable citizens of Canada a reasonable access to the ensured care of health, médicalement necessary and prepaid, without having to pay directly at the point service.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 4/4

A shared responsibility

Intelligibility: 1/5
Accuracy: 1/4

The federal government, the ten provinces and the three territories have roles paramount to play in the system of care of health in Canada.

Intelligibility: 2/5
Accuracy: 1/4

The federal government has the following responsibilities:

Intelligibility: 1/5
Accuracy: 1/4

· to establish and apply principles or standards on a national scale for the system of health within the framework of the Canadian Law to health (for further information, please consult the Web site of the Canadian Law on health);

Intelligibility: 3/5
Accuracy: 3/4

· to contribute to the financing of the provincial services of care of health, via payments of tax transfers;

Intelligibility: 2/5
Accuracy: 1/4

· to offer health services directly to certain groups such as ex-serviceman, the Canadians autochtones alive on the reserves, the military personnel, the prisoners of the federal penitentiaries and the royal Gendarmerie of Canada; and

Intelligibility: 3/5
Accuracy: 4/4

· to take up other duties relating to health like the protection of health, the prevention of the diseases and the promotion of health.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 1/4

The provincial and territorial governments have the following responsibilities:

Intelligibility: 1/5
Accuracy: 1/4

· to manage and exempt the assured health services (for further information on these schemes of health insurance, please consult the texts subjected by the provinces and the territories for the annual report on the Canadian Law on health);

Intelligibility: 3/5
Accuracy: 4/4

· to plan, finance and evaluate the services of hospital care, the medical departments and ancillary medical; and of

Intelligibility: 3/5
Accuracy: 4/4

· to manage certain aspects of the care on ordinance and public health.

Intelligibility: 2/5
Coherence: 1/5
Accuracy: 4/4

Coherence for text: 2/5

Summary of scoring:
Intelligibility:
Total score:
= 2+2+2+3+3+1+2+1+3+2+3+2+1+3+3+2 = 35
Average score:
= 35 total score / 16 sentences = 2.19
Coherence:
Total score:
= 1+2+2+1+2 = 8
Average score:
= 8 total score / 5 (4 paragraphs + text) = 1.6
Accuracy:
Total score:
= 1+1+2+4+4+1+1+1+3+1+4+1+1+4+4+4 = 37
Average score:
= 37 total score / 16 sentences = 2.31

Tester 3 - Health Care text - MT System 2
368 words
16 sentences
4 paragraphs

Care of health

Intelligibility: 2/5
Accuracy: 1/4

The financing of the Canadian system of health recovers in big part of the State.

Intelligibility: 2/5
Accuracy: 4/4

Our national regime of health insurance is a mosaic of thirteen provincial and territorial regimes of health insurance, bound{*connected*} by the respect for national principles established at federal level.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 2/4

Canadian Law on the health imposes criteria and conditions relatives on the assured{*insured*} health services and on the widened health services which provinces and territories have to respect to receive complete financial contribution in conformance with the Canadian Transfer in health and in social programs (CHST).

Intelligibility: 3/5
Accuracy: 3/4

Canadian Law on the health aims to assure{*insure*} to all the acceptable citizens of Canada a reasonable access to the assured{*insured*} care of health, medically necessary and prépayés, without having directly to pay to the point of service.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 4/4

A shared responsibility

Intelligibility: 1/5
Accuracy: 1/4

The federal government, ten provinces and three territories have roles original to play in the system of care of health in Canada.

Intelligibility: 2/5
Accuracy: 4/4

The federal government has the following responsibilities:

Intelligibility: 1/5
Accuracy: 1/4

- establish and apply principles or standards on a national scale for the system of health within the framework of the Canadian Law to the health (should you require further information, please consult the Web site of the Canadian Law on the health;

Intelligibility: 2/5
Accuracy: 3/4

- to help in the financing of the provincial services of care of health, by the intervention of payments of fiscal transfers;

Intelligibility: 2/5
Accuracy: 2/4

- to offer health services directly to certain groups such as the war veterans, the autochtonous Canadians living on reserves, military staff, prisoners of the federal penitentiaries and the royal Police force of Canada; and

Intelligibility: 2/5
Accuracy: 2/4

- to assume the other functions{*offices*} relative to the health as the protection of the health, the prevention of diseases and the promotion of the health.

Intelligibility: 3/5
Coherence: 1/5
Accuracy: 1/4

The provincial and territorial governments have the following responsibilities:

Intelligibility: 1/5
Accuracy: 1/4

- administer and distribute assured{*insured*} health services (should you require further information on these regimes of health insurance, please consult texts subjected with provinces and territories for the annual report on the Canadian Law on the health;

Intelligibility: 3/5
Accuracy: 4/4

- to plan, fork out and to estimate the performances of hospitable care, medical and ancillary medical services; and of

Intelligibility: 2/5
Accuracy: 4/4

- to administer certain aspects of care on prescription and health service.

Intelligibility: 2/5
Coherence: 3/5
Accuracy: 1/4

Coherence for text: 2/5

Summary of scoring:
Intelligibility:
Total score:

= 2+2+3+3+3+1+2+1+2+2+2+3+1+3+2+2 = 34
Average score:
= 34 total score / 16 sentences = 2.13

Coherence:

Total score:
= 2+2+1+3+2 = 10
Average score:
= 10 total score / 5 (4 paragraphs + text) = 2.00

Accuracy:

Total score:
= 1+4+2+3+4+1+4+1+3+2+2+1+1+4+4+1 = 38
Average score:
= 38 total score / 16 sentences = 2.38

Tester 4 - Corporate Overview – National Research Council - MT System 1
400 words
25 sentences
10 paragraphs

Overall picture of the CNRC

Intelligibility: 1/5
Accuracy: 1/4

Structure

Intelligibility: 1/5
Accuracy: 1/4

The CNRC gathers 19 national institutes and programs covering a multitude of fields and a large variety of services.

Intelligibility: 2/5
Accuracy: 2/4

The CNRC has offices in all the provinces of Canada that and plays a part determining by feeding the innovation within the communities.

Intelligibility: 3/5
Coherence: 3/5
Accuracy: 2/4

The institutes and programs of the CNRC are structured in three (3) great sectors:

Intelligibility: 2/5
Accuracy: 2/4

§ the genius and research Program in physical sciences

Intelligibility: 4/5
Accuracy: 4/4

§ the research Programme in life sciences and information technologies

Intelligibility: 2/5
Accuracy: 2/4

§ the technological and industrial Support

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 1/4

The genius and research Program in physical sciences includes/understands the four institutes of technology of manufacture (CI-CNRC, ITFI-CNRC, ITPCE-CNRC and IMI-CNRC) and four institutes (IRA-CNRC, IRC-CNRC, ITO-CNRC and IHA-CNRC).

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 4/4

The research Program Life sciences and information technologies, wallet which gathers the five institutes of biotechnology (IRB-CNRC, ISB-CNRC, IBP-CNRC, IBD-CNRC, IBM-CNRC), the initiative into genomic and health, the ISSM-CNRC, the INN-NRC, the IENM-CNRC and institutes of the sector of communication and information technologies, the ISM-CNRC and the ITI-CNRC.

Intelligibility: 3/5
Coherence: 3/5
Accuracy: 4/4

The technological and industrial Support gathers various services intended for industry, in particular the Canadian Institute of scientific and technical Information and the Supplementary programme to the industrial research of the CNRC, which stresses the needs for small and medium-sized undertakings.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 3/4

Control

Intelligibility: 4/5
Accuracy: 4/4

The CNRC is an organization of the government of Canada.

Intelligibility: 2/5
Accuracy: 1/4

It returns accounts to the Parliament by the means of the Minister of Industry and is managed by a council of 22 members chosen among his customers.

Intelligibility: 2/5
Accuracy: 3/4

To know the composition of the board of directors, please activate the bond below or cliquer the mitre " Board of directors " on the bar of navigation, on the left.

Intelligibility: 3/5

Coherence: 3/5
Accuracy: 3/4

Of to know more:

Intelligibility: 3/5
Accuracy: 2/4

Board of directors

Intelligibility: 1/5
Accuracy: 1/4

List current members of the Board of directors of the CNRC.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 1/4

Flow chart of the CNRC

Intelligibility: 1/5
Accuracy: 1/4

The organisational structure of the national Council of research (institutes, directions and programs), as well as the team of management.

Intelligibility: 1/5
Coherence: 2/5
Accuracy: 3/4

Biography of the frameworks

Intelligibility: 4/5
Accuracy: 4/4

Learn more on the top management from the CNRC.

Intelligibility: 2/5
Accuracy: 2/4

You will find there a biography brief of the president, vice-presidents and of the secretary-general of the CNRC, as well as the list of the general managers of the institutes, programs and general directions of the Council.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 2/4

Capsules

Intelligibility: 3/5
Accuracy: 2/4

Learn more on the CNRC, in particular its manpower, its budget, the site of its services in Canada, its achievements historical, the Law on the CNRC and the files of the Council.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 3/4

Coherence of text: 2/5

Summary of scoring:
Intelligibility:
Total score:
= 1+1+2+3+2+4+2+2+2+3+3+4+2+2+3+3+1+2+1+1+4+2+3+3+2 = 58
Average score:
= 58 total score / 25 sentences = 2.32
Coherence:
Total score:
= 3+2+2+3+2+3+2+2+2+2+2 = 25
Average score:
= 25 total score / 11 (10 paragraphs + text) = 2.27
Accuracy:
Total score:
= 1+1+2+2+2+4+2+1+4+4+3+4+1+3+3+2+1+1+1+3+4+2+2+2+3 = 58
Average score:
= 58 total score / 25 sentences = 2.32

Tester 4 - NRC text - MT System 2
403 words
25 sentences
10 paragraphs

General view of the CNRC

Intelligibility: 1/5
Accuracy: 2/4

Structure

Intelligibility: 1/5
Accuracy: 1/4

The CNRC regroups 19 institutes and national programs covering a multitude of domains and a big variety of services.

Intelligibility: 2/5
Accuracy: 3/4

The CNRC has desks in all the provinces of Canada which and plays a determining role by feeding innovation within communities.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 3/4

Institutes and programs of the CNRC are structured in three (3) big sectors:

Intelligibility: 2/5
Accuracy: 2/4

- § the Program of search in physical sciences and the genius

Intelligibility: 3/5
Accuracy: 4/4

- § the Program of search in life sciences and information technologies

Intelligibility: 3/5
Accuracy: 2/4

- § technological and industrial Support

Intelligibility: 1/5
Coherence: 2/5
Accuracy: 1/4

The Program of search in physical sciences and the genius understands{*includes*} the four institutes of technology of manufacture (CI-CNRC, ITFI-CNRC, ITPCE-CNRC AND IMI-CNRC) and four institutes (IRA-CNRC, IRC-CNRC, ITO-CNRC AND IHA-CNRC).

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 2/4

The Program of search Life sciences and information technologies, wallet which regroups the five institutes of biotechnology (IRB-CNRC, ISB-CNRC, IBP-CNRC, IBD-CNRC, IBM-CNRC), the initiative in génomique and in health, the ISSM-CNRC, the INN-CNRC, the IENM-CNRC and the institutes of the sector of information technologies and communications, ISM-CNRC and ITI-CNRC.

Intelligibility: 3/5
Coherence: 3/5
Accuracy: 3/4

Technological and industrial Support collects different services intended for the industry, notably the Canadian Institute of the scientific and technical Information and the Program of help{*assistant*} in the industrial search for the CNRC, which emphasizes the necessities of small and medium-sized companies.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 3/4

State control

Intelligibility: 2/5
Accuracy: 3/4

The CNRC is a body of the government of Canada.

Intelligibility: 1/5
Accuracy: 2/4

It{*he*} returns accounts to the Parliament with the aid of Secretary of Commerce and is administered by an advice{*council*} of 22 members chosen among the clientele.

Intelligibility: 3/5
Accuracy: 3/4

To know the composition of the board of directors, please activate link below or click the thumb index "Board of directors" the bar of navigation, to the left.

Intelligibility: 2/5
Coherence: 3/5
Accuracy: 2/4

To know more about it:

Intelligibility: 1/5
Accuracy: 1/4

' Council dadministration

Intelligibility: 2/5
Accuracy: 4/4

List current members of the Board of directors of the CNRC.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 1/4

Organization chart of the CNRC

Intelligibility: 1/5
Accuracy: 1/4

The organizational structure of the national council of searches (institutes, directions{*managements*} and programs), as well as the team of management.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 2/4

Biography of the executives

Intelligibility: 2/5
Accuracy: 1/4

Learn there more on the high direction{*management*} of the CNRC.

Intelligibility: 2/5

Accuracy: 2/4

You will find there a brief biography of the president, the vice-presidents and the general secretary of the CNRC, as well as the list of the general managers of institutes, programs and head offices of the council.

Intelligibility: 2/5
Coherence: 2/5
Accuracy: 1/4

Capsules

Intelligibility: 3/5
Accuracy: 2/4

Learn there more on the CNRC, notably the actual, its budget, the place of its services in Canada, its historic realizations, Law on the CNRC and the archives of the council.

Intelligibility: 3/5
Coherence: 2/5
Accuracy: 2/4

Coherence of text: 2/5

Summary of scoring:
Intelligibility:
Total score:
= 1+1+2+2+2+3+3+1+3+3+3+2+1+3+2+1+2+2+1+2+2+2+2+3+3 = 52
Average score:
= 52 total score / 25 sentences = 2.08
Coherence:
Total score:
= 2+2+2+3+2+3+2+2+2+2+2 = 24
Average score:
= 24 total score / 11 (10 paragraphs + text) = 2.18
Accuracy:
Total score:
= 2+1+3+3+2+4+2+1+2+3+3+3+2+3+2+1+4+1+1+2+1+2+1+2+2 = 53
Average score:
= 53 total score / 25 sentences = 2.12

Appendix E Post-edited Raw Output

1. SARS text - Systran

Questions and Answers - Severe Acute Respiratory Syndrome (SARS)
Health in the workplace

Q. How does Health Canada provide support to Government of Canada workers in respect with SARS?

A. The role of Health Canada's Workplace Health and Public Safety Programme (WHPSP) consists of advising Government of Canada workers in connection with issues of occupational health and safety. Workplace Health and Public Safety Programme health professionals were on hand at the Pearson, Vancouver and Dorval airports to organize information meetings for Government of Canada workers. They continue to organize information meetings for employers and employees in work environments everywhere in the country, and to answer employers' daily requests for information.
The WHPSP issued advisories and members of its personnel have spoken to or met with representatives of a number of federal departments in order
to provide them with information and advice on health and safety measures considered appropriate for their employees.

End of first half
Total time taken to post-edit first half: 11 minutes

The representatives of WHPSP attended a videoconference with employees
working in seven missions located abroad in the affected areas in order
to provide them with information on the protection of the respiratory
tract and to discuss other general health concerns about SARS.

Q. Why doesn't Health Canada recommend the use of masks by Government of Canada workers such as customs agents at the Pearson and Vancouver airports?

A. The role of Health Canada's WHPSP consists of providing
advice to Government of Canada workers on issues of occupational health and safety. Health Canada believes that federal government employees
are not likely to become sick, for the contacts of employees with
passengers or customers are of relatively short duration and
are not regarded as being close contact. Consequently, Health Canada considers that the use of the masks is not necessary. Employers can use this information and advice to determine health and safety measures they consider suitable for their employees.

End of second half
Total time taken to post-edit second half: 8 minutes

Total time taken to post-edit text: 19 minutes

1. SARS text - ReversoPro

Questions and answers - austere pointed respiratory Syndrome (SARS)
in the workplace.

Q. How does Health Canada provide support to federal civil servants in SARS related issues?

A. The role of Health Canada's Workplace Health and Public Safety Programme (WHPSP) consists of advising federal employees on occupational health and safety issues. Health Canada's WHPSP health professionals on site at the Pearson, Vancouver and Dorval airports organized information sessions for federal employees. They continue to organize information sessions of information for employers and employees in workplaces across the country and to answer the daily information requests received by employers. The WHPSP has issued advisories and staff members have spoken to or met with representatives of a number of federal departments to provide them with information and advice about the health and safety measures considered appropriate for their employees.

End of first half
Total time taken to post-edit first half: 7 minutes

WHPSP representatives attended a videoconference with employees working in seven missions abroad in the affected regions to provide them with information on respiratory tract protection and to discuss other general health preoccupations about SARS.

Q. Why doesn't Health Canada recommend the use of masks by federal civil servants such as customs officers in the Pearson and Vancouver airports?

A. The role of Health Canada's WHPSP consists of providing advice to federal government employees on occupational health and safety matters. Health Canada believes that federal government employees are not at risk of becoming ill because their contacts with passengers and customers are of relatively short duration and are not considered to be close contact. Consequently, Health Canada considers that the use of masks is not necessary. Employers may use this information and advice to determine the health and safety measures they consider necessary for their employees.

End of second half
Total time taken to post-edit second half: 10 minutes

Total time taken to post-edit text: 17 minutes

Tester 1 - Canadian Dairy Commission Text - MT System 1
(raw output = 341 words)

History

Since the appointment of the first Dominion Dairy Commissioner in 1890, the federal government has played an active role in policies and programs development and implementation that supports the milk industry.

Initial Programs

The initial programs included:

- a refrigerated railway car service for butter (implemented in 1895);
- cool rooms for cheese curing (as of 1902);

- tests for dairy cows (from 1902 onward); and
- a classification system for butter and cheese intended for export (in 1923).

Subsequently, the Government of Canada introduced a temporary subsidy on cheese and butter. During the '40s and '50s, it launched programs to support prices, export surplus products, and limit imports. These programs continued following the creation of the Agricultural Stabilization Board in 1958.

Although it offered the necessary structure to stabilize prices, the Agricultural Stabilization Board was not able to tackle two major problems which seriously affected the dairy sector:

1) the lack of coordination between the federal and provincial policies, and
2) the absence of an effective mechanism to control dairy production.

Such a mechanism proved to be necessary to ensure reasonable support prices for storable dairy products in order to underline producer returns. New mechanisms were also required to control costly surplus production. In 1963, the federal government organized the Canadian Dairy Conference, which led to the creation of the Canadian Dairy Advisory Committee that year. In its 1965 final report, the Committee recommended the creation of the Canadian Dairy Commission.

The Canadian Dairy Commission (CDC) was established by the proclamation of the Canadian Dairy Commission Act on October 31, 1966. The creation of this Crown Corporation confirmed the federal government's continuing commitment to ensure a strong and viable dairy industry in Canada.

Total post-editing time: 21 minutes

Average words per minute:
= number of words in raw output / total post-editing time
= 341 / 21
= 16.24 words per minute

Tester 1 - Canadian Dairy Commission text - MT System 2
(raw output = 333 words)

History

Since the appointment of the first Dominion Dairy Commissioner in 1890, the federal government has played an active role in the development and application of policies and programs to support dairy farming.

Initial Programs

Initial programs included a refrigerated railway car service for butter (1895), cooled rooms to cure cheese (1902), dairy cow testing (1902), and a classification system for butter and cheese intended for export (1923). Subsequently, in 1935, the Government of Canada adopted a temporary subsidy for cheese and butter. During the 1940s and 1950s, it launched programs to support prices, export superfluous products, and limit imports. This program continued after the creation of the Agricultural Stabilization Board in 1958.
Although it provided the necessary structure to stabilize prices, the Agricultural Stabilization Board was not able to tackle two major problems which seriously affected the dairy sector:

1) the lack of coordination between federal and provincial policies, and
2) the absence of an effective mechanism to control dairy production.

Such a mechanism turned out to be necessary to ensure reasonable support prices for stored dairy products to strengthen the producers' revenues. New mechanisms were also required to control the expensive production surpluses. In 1963, the federal government organized the Canadian Dairy Conference, which led to the creation of the Canadian Dairy Advisory Committee that same year. The final report of this Committee, which was presented in 1965, recommended the creation of the Canadian Dairy Commission.

The Canadian Dairy Commission (CDC) was established by the declaration of Canadian Dairy Commission Act on October 31, 1966. The creation of this Crown Corporation, confirmed the federal government's commitment to ensuring the existence of a strong and viable dairy farming industry in Canada.

Total post-editing time: 15 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 333 / 15
= 22.27

Tester 2 - Income Tax text - MT System 1
(raw output = 368 words)

Do you have to file an income tax return?

You must file an income tax return for 2003 if you are in one of the following situations:

· you have to pay tax in 2003;

· we asked you to file a return;

· you disposed of property in 2003 (for example, you sold real estate or shares) or you realized a taxable capital gain (for example, a mutual fund or a trust attributed amounts to you, or you must report a capital gains reserve which you claimed in 2002);

· you must repay some or all of the Old Age Security or Employment Insurance benefits you received. See line 235 – Social Benefits Repayment;

· you did not repay all of the amounts which you withdrew from your registered retirement savings plan within the framework of Home Buyers' Plan or the Lifelong Learning Plan. For more information, consult guide RC4135 – Home Buyers' Plan or guide RC4112 – Lifelong Learning Plan (LLP);

· you have to contribute to the Canada Pension Plan (CPP) because, in 2003, the total of your net self-employment income pensionable employment income exceeds $3,500. See line 222 - Deduction for CPP or QPP contributions on self-employment and other earnings.

Note

Even if you are not in one of these situations, you can still file an income tax return. To receive certain benefits and credits, you must file an income tax return. Read Benefits of filing a tax return.

What is new

We will pay you compound daily interest on your tax refund for 2003 on the latest of the following dates:

· May 31, 2004;

· the 31st day after you file your tax return;

· the day following after you overpaid your taxes.

Important dates

In general, you must file your tax return for 2003 with us on or before April 30, 2004.

Total post-editing time: 15 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 368 / 15
= 24.53

Tester 2 - Income Tax text - MT System 2
(raw output = 392 words)

Do you have to file a tax return?

You have to file a tax return for 2003 if you are in one of the following situations:

- you have to pay tax in 2003;
- we asked you to file a tax return;
- you disposed of property in 2003 (for example, you sold real estate or shares) or you realized a taxable capital gain (for example, a mutual fund or a trust attributed amounts to you, or you have to report a capital gains reserve you claimed in 2002);
- you have to pay back some or all of your Old Age Security or Employment Insurance benefits. See line 235 – Social benefits repayment;
- you did not repay all of the amounts which you withdrew from your registered retirement savings plan within the framework of the Home Buyers' Plan or the Lifelong Learning Plan. To find out more, consult guide RC4135 – Home Buyers' Plan or guide RC4112 – Lifelong Learning Plan (LLP);
- you have to contribute to the Canada Pension Plan because, in 2003, the total of your net self-employment income and pensionable employment income exceeds $3,500. See line 222 - Deduction for CPP or QPP contributions on self-employment and other earnings.

Note

Even though you are not in one of these situations, you can still file a tax return. To receive certain benefits and credits, you have to file a tax return. Read Advantages of filing a tax return.

What's new

We will pay you compound daily interest on your tax refund for 2003 on the latest of the following dates:

- May 31, 2004;
- the 31st day after you file your tax return;
- the day after you overpaid your taxes.

Important dates

Generally, you have to file your tax return for 2003 on or before April 30, 2004.

Total post-editing time: 10 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 392 words / 10 minutes
= 39.20

Tester 3 - Health Care text - MT System 1
(raw output = 359 words)

Health Care

The financing of the Canadian system of health is mainly provided by the state. Our national health insurance program is a mosaic of thirteen provincial and territorial health insurance programs, linked through respect of national principles established at the federal level.

The Canada Health Act imposes criteria and conditions relating to insured health services and extended health services that the provinces and the territories must respect in order to receive the full financial contribution under the Canada Health and Social Transfer (CHST). The Canada Health Act aims to ensure that all eligible Canadian residents have reasonable access to medically necessary insured health care on a prepaid basis, without having to pay directly at the point of service.

A shared responsibility

The federal government, the ten provinces and the three territories have fundamental roles to play in Canada's health care system. The federal government has the following responsibilities:

· to establish and apply principles or standards on a national scale for the health system within the framework of the *Canada Health Act* (for further information, please consult the *Canada Health Act* website);

· to contribute to the financing of provincial health care services via fiscal transfers;

· to offer health services directly to certain groups such as war veterans, native Canadians living on reserves, military personnel, inmates of federal penitentiaries and the Royal Canadian Mounted Police; and

· to fulfill other health-related duties such as health protection, disease prevention and health promotion.

The provincial and territorial governments have the following responsibilities:

· to manage and deliver insured health services (for further information on these health insurance programs, please consult the provincial and territorial submissions to the *Canada Health Act* Annual Report);

· to plan, finance and evaluate the provision of hospital care, physiscian and allied health care services; and

· to manage certain aspects of prescription care and public health.

Total post-editing time: 13 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 359 / 13
= 27.62

Tester 3 - Health Care text - MT System 2
(raw output = 368 words)

Health Care

The Canadian health care system is predominately funded by the State.

Our national health insurance program is a mosaic of thirteen provincial and territorial health insurance programs, connected by respect for national principles established at the federal level.

The *Canada Health Act* imposes criteria and conditions relative to insured health services and extended health services, which provinces and territories have to respect in order to receive the full federal cash contribution in accordance with the Canada Health and Social Transfer (CHST).

The *Canada Health Act* aims to guarantee all eligible Canadian citizens of Canada reasonable access to medically necessary insured health care on a prepaid basis, without having to pay directly at the point of service.

A shared responsibility

The federal government, the ten provinces and the three territories have essential roles to play in the Canada's health care system.

The federal government has the following responsibilities:

- to establish and administer principles or standards on a national scale for the health care system within the framework of the *Canada Health Act* (should you require further information, please consult the *Canada Health Act* website);

- to help in the financing of provincial health care services through fiscal transfer payments;

- to offer health services directly to certain groups such as war veterans, native Canadians living on reserves, military staff, inmates in federal penitentiaries and the Royal Canadian Mounted Police; and

- to assume other functions related to health protection, disease prevention and health promotion.

The provincial and territorial governments have the following responsibilities:

- to administer and deliver insured health services (should you require further information on these health insurance programs, please consult the provincial and territorial submissions to the *Canada Health Act* Annual Report);

- to plan, finance and evaluate the provision of hospital care, physician and allied health care services; and

- to administer certain aspects of prescription care and public health.

Total post-editing time: 17 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 368 / 17
= 21.65

Tester 4 - NRC Text - MT System 1
(raw output = 400 words)

Overview of the NRC

Structure

The NRC comprises 19 national institutes and programs covering a multitude of fields and a wide variety of services. The NRC has offices in all the provinces of Canada and plays an important role by supporting innovation within communities.

The institutes and programs of the NRC are organized into three (3) main areas:

§ the Physical Sciences and Engineering Research Program
§ the Life Sciences and Information Technology Research Program
§ Technological and Industrial Support

The Physical Sciences and Engineering Research Program includes the four institutes of technology of manufacture (CI-NRC, ITFI-NRC, ITPCE-NRC and IMI-NRC) and four additional institutes (IRA-NRC, IRC-NRC, ITO-NRC and IHA-NRC).

The Life Sciences and Information Technology Research Program is a portfolio that includes the five institutes of biotechnology (IRB-NRC, ISB-NRC, IBP-NRC, IBD-NRC, IBM-NRC), the genomics and health initiative, the ISSM-NRC, the INN-NRC, the IENM-NRC and institutes
from the information technology and communications sector, the ISM-NRC and the ITI-NRC.

Technological and Industrial Support includes various services intended for industry, in particular the Canadian Institute of Scientific and Technical Information and the NRC's Industrial Research Assistance Program, which focuses on the needs of small and medium-sized businesses.

Control

The NRC is an agency of the government of Canada. It reports to Parliament through the Minister of Industry and is managed by a council of 22 members that are chosen from its clients.

For a list of the Council members, please activate the link below or click the "Council Members" tab on the navigation bar on the left.

For more information:

Council members

A list of current NRC Council members.

NRC flow chart

The organizational structure of the National Research Council (institutes, branches and programs), as well as the management team.

Executive biographies

Learn more about the upper management of the NRC. You will find a brief biography of the president, vice-presidents and the secretary-general of the NRC, as well as the list of the general
managers of the institutes, branches and general directions of the Council.

Capsules

Learn more about the NRC, in particular its workforce, its budget, the location of its services in Canada, its historical achievements, the NRC Act and the Council archives.

Total time to post-edit: 15 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 400 / 15
= 26.67

Tester 4 - NRC Text - MT System 2
(raw output = 403 words)

General overview of the NRC

Structure

The NRC comprises 19 national institutes and programs covering several domains and a wide variety of services. The NRC has offices in all the provinces of Canada and plays a determining role by supporting innovation within communities.

The NRC's institutes and programs are organized into three (3) main groups:

- § the Physical Science and Engineering Research Program
- § the Life Sciences and Information Technology Research Program
- § Technological and Industrial Support

The Physical Science and Engineering Research Program is made up of the four manufacturing technology institutes (CI-NRC, ITFI-NRC, ITPCE-NRC and IMI-NRC) in addition to the following four institutes: IRA-NRC, IRC-NRC, ITO-NRC and IHA-NRC.

The Life Sciences and Information Technology Research Program is a portfolio that is made up of the five institutes of biotechnology (IRB-NRC, ISB-NRC, IBP-NRC, IBD-NRC, IBM-NRC), the genomics and health initiative, the ISSM-NRC, the INN-NRC, the IENM-NRC and the institutes of the information technology and communications sector, the ISM-NRC and the ITI-NRC.

Technological and Industrial Support refers to various industry services, notably the Canadian Institute of Scientific and Technical Information and the NRC's Industrial Research Assistance Program which focuses on the needs of small and medium-sized companies.

State control

The NRC is an agency of the government of Canada. It reports to Parliament through the Minister of Industry and is administered by acouncil of 22 members chosen among the clientele. To see a list of the board of directors, please activate the link below or click on the "Board of directors" tab on the navigation barat the left.

For more information:

NRC Council
A list of the current members of the Board of directors of the NRC.

Organization chart of the NRC
The organizational structure of the National Research Council (institutes, branches and programs), as well as the management team.

Executive Biographies
Learn more about the NRC's upper management. You will find a brief biography of the president, the vice-presidents and the general secretary of the NRC, as well as the list of the general managers of institutes, branches and head offices of the Council.

Capsules
Learn more about the NRC, notably its workforce, its budget, the location of its services in Canada, its past accomplishments, the NRC Act and the archives of the council.

Total post-editing time: 25 minutes

Average words per minute:
= total words in raw output / total post-editing time
= 403 / 25
= 16.12

Appendix F Instructions for Testers

General Instructions
2003-03-22

You have been asked to evaluate and revise (post-edit) the output of two MT systems as part of the testing of an evaluation methodology developed for my M.A. thesis. The purpose of this evaluation methodology is not to determine the absolute quality of the MT output, but rather to see if it is possible for you to determine which of the two MT systems used produces better quality output, as reflected by the scoring of the raw output and the time taken to post-edit it.

In total, four testers will take part in this experiment. Each of you will be given a diskette containing the data required to apply the methodology to one short text of approximately 350 words. Each tester will evaluate a different text. Each diskette contains three folders as follows: MT System1 raw output, MT System2 raw output and texts for post-editing.

The testing will take place in two parts. In the first part you will be asked to evaluate the raw output of the two MT systems for intelligibility, coherence, and accuracy, in that order. Both the MT System1 and MT System2 folders contain three documents: one for each aspect of the scoring, appropriately labelled *intelligibility*, *coherence*, or *accuracy*. Please refer to the attached detailed instructions for scoring scales.

In the second part you will be asked to post-edit the two translations of the text. Please note that Testers 1 and 2 should post-edit the output of MT System1 first, followed by the output of MT System2. In contrast, Testers 3 and 4 should edit the output of MT System2 first, followed by the output of MT System1.

I. Scoring of raw output

Please apply the scoring metrics for intelligibility, coherence, and accuracy (in that order) to the documents in the folders labelled MT System1 and MT System2 using the detailed instructions attached.

II. Post-editing

Please post-edit the texts in the folder labelled Post-editing in the order outlined above (i.e., Testers 1 and 2 post-edit MT System1 output first, while Testers 3 and 4 post-edit MT System2 output first). You need only keep track of the time taken to the nearest minute for each document. Please note the time taken for each at bottom of the document in the space indicated.

Please apply Canadian spelling and usage (including typographical conventions). Change only what is necessary to make the text accurate, intelligible and coherent. Do not make changes or revisions for style or to reflect personal preferences.

Thank you!

Appendix F

Instructions for Testing
Details of Scoring Scales Applied to Raw Output
2004-03-22

Note that sentence length and therefore number of errors may vary. However, the seriousness of the errors can be seen to compensate for this. Accordingly, the minimum/maximum numbers given are guidelines rather than hard and fast rules.

Step 1

Please rate each sentence or segment of your texts for intelligibility using the following 5-point scale. You need only assign a number score in the space indicated. Do not award partial points for scoring (i.e. 1.5 or 3.5). Note that the more intelligible or readable the sentence, the lower the score. Accuracy and coherence should not be considered when assigning intelligibility scores. Table 1 contains examples of intelligibility errors and how they affect the scores.

Intelligibility 5-point scale:

1. Completely intelligible: no rewriting or editing required.
2. Intelligible: but some problems with usage or word order (up to 3 errors in a simple sentence).
3. Adequately intelligible: basic meaning can be understood but some details are not clear and contains a maximum of 5 errors of usage or word order or other grammatical errors.
4. Unintelligible: reader can only guess at meaning: many and more serious grammatical errors than for category 3 (i.e. more than 5 errors of usage and word order, but not more than 10).
5. Completely unintelligible: more than 10 errors of usage and word order or serious grammatical errors that make sentence unintelligible.

Table 1 - Examples of different types of intelligibility-related errors

Intelligibility category:	**Example**	**Explanation**
1. completely intelligible	Consequently, Health Canada considers that the custom{*usage*} of masks is not necessary.	There are no usage or word order errors in this sentence. **Custom/usage** is a meaning error and is taken into account in the accuracy scoring.
2. intelligible	**The** employers can use this information and	The articles **the** and

	these councils in order to determine measurements of health and safety which they consider suitable for their employees.	**these** need to be deleted. (2 errors) **Councils** and **measurements** are meaning errors and taken into account in the accuracy scoring.
3. adequately intelligible	Q. Why Health Canada **does not he** recommend the custom{*usage*} of masks by **the** federal civil servants such as **the** customs officers in **airports Pearson** and **of Vancouver**?	Word order errors include **does not he** and **airports Pearson** and **Vancouver**. The articles **the** (federal…) and **the** (customs…) as well as the preposition **of** need to be deleted. (5 errors) **Custom/usage** is a meaning error and taken into account in the accuracy scoring.
4. unintelligible	Health Canada believes that **the** employees **of the federal government do not risk to become_patients**, because **the contacts of the employees** with **the** passengers or **the** customers are of relatively short duration and are not considered as narrow.	The articles **the** (employees), **the** (passengers) and **the** (customers) need to be deleted. Word order errors include **of the federal government** and **the contacts of the employees**. Verb tense and form errors include **do not risk** and **to become**. (7 errors) **As narrow** is a meaning error and taken into account in the accuracy scoring.
5. completely unintelligible	How Santé Canada **does it** support **the** federal **civils** servant **in what** milked **with the** SRAS?	Word order errors include **does it**. The articles **the** (federal) and **the** (SRAS) need to be deleted. The **s** should be deleted in **civils** and should be added to **servant**. **In what** is grammatical error that needs to be reformulated as **with (the SRAS)**. **Santé Canada**, **SRAS** and **milked** are

		meaning errors and taken into account in the accuracy scoring.

Step 2

Please rate each paragraph or section of your texts for coherence using the following 5-point scale. Also, please rate the text as a whole for coherence using the same 5-point scale. You need only assign a number score in the space indicated. Do not award partial points (i.e. 1.5 or 3.5). Note that the more accurate the paragraph or segment, the lower the score. Accuracy and intelligibility at the sentence level should not be considered when assigning coherence scores.

Coherence 5-point scale

1. Completely coherent: passage (i.e. paragraph or text) is logical and consistent. No editing required for coherence.

2. Coherent: passage is logical but contains some problems, especially with pronoun reference. No minimum or maximum number of errors has been established. Paragraphs and texts vary in length so that the number of errors that result in the passage being incoherent will vary.

3. Adequately coherent: overall meaning can be understood but some sentences (i.e. up to 50%) in passage are not logical.

4. Incoherent: reader can only guess at overall meaning: many (i.e. more than 50% of the sentences in passage are not logical.

5. Completely incoherent: meaning of paragraph or document cannot be detected.

Step 3

Please rate each sentence or section of your texts for accuracy using the following 4-point scale. You need only assign a number score in the space indicated. Do not award partial points for scoring (i.e. 1.5 or 3.5). Note that the more accurate the sentence, the lower the score. Intelligibility and coherence should not be considered when assigning accuracy scores. Table 2 contains examples of meaning (accuracy) errors and how they affect the scores.

Accuracy 4-point scale

1. Completely accurate: no meaning errors.

2. Accurate: minor meaning errors (no more than 2), especially words/terms not in dictionary.

3. Adequately accurate: minor meaning errors (2 to 5) including minor omissions and verb tense errors. Errors do not fundamentally change meaning.

4. Inaccurate: major meaning error including major omission or more than 5 minor meaning errors. Error changes fundamental meaning.

Table 2 - Examples of different types of meaning errors

Error type	**Example**	**Explanation**
Minor meaning error:	ST: fournir des **conseils** au fonctionnaires fédéraux MT: providing **councils** to the federal civils servant	The noun **conseils** has been translated as **councils** instead of **advice**. The meaning can still be understood.
Major meaning error:	ST: en ce qui **a trait** au SRAS MT: in what **milked** with the SRAS	The verb **traiter** (to deal with) has been incorrectly parsed as **traire** (to milk). The meaning has been fundamentally changed.

Appendix G Questionnaires

Evaluation procedure questionnaire
To be filled out following the testing
2004-03-22

In order to help evaluate this methodology and any potential improvements, please answer the following questions:

1) Did you find the evaluation metrics (intelligibility, coherence, accuracy) straightforward to apply? If not, what difficulties did you encounter?

2) Was the post-editing process relatively straightforward? If not, what problems did you encounter?

3) Do you have any suggestions for modifications to either the evaluation metrics or post-editing process?

4) If you were trying to choose between two MT systems for use in your work, would you find the methodology proposed in this experiment to be useful?

5) Do you have any additional comments about this experiment or the methodology used?

Appendix G

Tester 1 Questionnaire
Completed following the testing
2004-03-22

In order to help evaluate this methodology and any potential improvements, please answer the following questions:

1) Did you find the evaluation metrics (intelligibility, coherence, accuracy) straightforward to apply? If not, what difficulties did you encounter?

Yes, but sometimes I may have mixed them up (esp. the last 2).

2) Was the post-editing process relatively straightforward? If not, what problems did you encounter?

Yes.

3) Do you have any suggestions for modifications to either the evaluation metrics or post-editing process?

It may be better to not do all tests in one sitting. I found it was easier to do the 2nd texts (which required more editing) because I already knew the source from doing the 1st post-editing.

4) If you were trying to choose between two MT systems for use in your work, would you find the methodology proposed in this experiment to be useful?

Yes—it was clear which system I would prefer esp. regarding metrics 1 and 3.
(MT System 1).

5) Do you have any additional comments about this experiment or the methodology used?

Appendix G

Tester 2 Questionnaire
Completed following the testing
2004-03-22

In order to help evaluate this methodology and any potential improvements, please answer the following questions:

1) Did you find the evaluation metrics (intelligibility, coherence, accuracy) straightforward to apply? If not, what difficulties did you encounter?

Pretty straightforward. They took some time to get used to, but were much easier for the second text. Biggest problems were assessing how serious some of the errors really were.

2) Was the post-editing process relatively straightforward? If not, what problems did you encounter?

Yes, it was straightforward. My biggest problem was terminology. It would have been much harder if you hadn't provided the "human translation."

3) Do you have any suggestions for modifications to either the evaluation metrics or post-editing process?

4) If you were trying to choose between two MT systems for use in your work, would you find the methodology proposed in this experiment to be useful?

Yes (and I would choose MT 2 simply because it provided additional suggestions.)

5) Do you have any additional comments about this experiment or the methodology used?

Appendix G

Tester 3 Questionnaire
Completed following the testing
2004-03-22

In order to help evaluate this methodology and any potential improvements, please answer the following questions:

1) Did you find the evaluation metrics (intelligibility, coherence, accuracy) straightforward to apply? If not, what difficulties did you encounter?

Intelligibility and accuracy were very straightforward, but coherence was more difficult. Examples may have helped. I gave relatively good coherence scores, but I wasn't quite certain what I was looking for. Perhaps it would have been more obvious had it actually been a real problem in the texts.

2) Was the post-editing process relatively straightforward? If not, what problems did you encounter?

Yes.

3) Do you have any suggestions for modifications to either the evaluation metrics or post-editing process?

No.

4) If you were trying to choose between two MT systems for use in your work, would you find the methodology proposed in this experiment to be useful?

Yes, from a tester's perspective. It would also depend on the effort and cost involved in preparing for the test.

5) Do you have any additional comments about this experiment or the methodology used?

In general, the instructions are quite clear. It is easy to distinguish between the categories (i.e. I didn't have to hesitate very often when trying to choose a score). That bodes well for consistent application. I'm still not sure how I would react to a less coherent text.

Appendix G

Tester 4 Questionnaire
Completed following the testing
2004-03-22

In order to help evaluate this methodology and any potential improvements, please answer the following questions:

1) Did you find the evaluation metrics (intelligibility, coherence, accuracy) straightforward to apply? If not, what difficulties did you encounter?

A clear definition for what is meant by intelligibility, coherence, and accuracy.

2) Was the post-editing process relatively straightforward? If not, what problems did you encounter?

Yes.

3) Do you have any suggestions for modifications to either the evaluation metrics or post-editing process?

No.

4) If you were trying to choose between two MT systems for use in your work, would you find the methodology proposed in this experiment to be useful?

Yes, although the number of errors the systems make might be more important than the types.

5) Do you have any additional comments about this experiment or the methodology used?