



uOttawa

L'Université canadienne
Canada's university

**FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES**



**FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES**

Matthew Thomas Fifer

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A. (Translation)

GRADE / DEGRÉ

School of Translation and Interpretation

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

**The Fuzzy Factor: An Empirical Investigation of Fuzzy Matching in the Context of Translation
Memory Systems**

TITRE DE LA THÈSE / TITLE OF THESIS

Lynne Bowker

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

A. Francoeur

J. Quirion

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

**The Fuzzy Factor:
An Empirical Investigation of Fuzzy Matching
in the Context of Translation Memory Systems**

by / par

Matthew Fifer

School of Translation and Interpretation / École de traduction et d'interprétation

University of Ottawa / L'Université d'Ottawa

Under the supervision of / Sous la direction de

Lynne Bowker, Ph.D

Thesis submitted to the / Thèse soumise à la

**Faculty of Graduate and Postdoctoral Studies of the University of Ottawa /
Faculté des études supérieures et postdoctorales de l'Université d'Ottawa**

**In partial fulfillment of the requirements for the degree M.A. (Translation) /
Dans le cadre des exigences du programme de maîtrise en traduction**

© Matthew Fifer, Ottawa, Ontario, Canada, 2007.



Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 978-0-494-34069-1

Our file Notre référence

ISBN: 978-0-494-34069-1

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.


Canada

Table of Contents

List of Tables	vi
List of Figures	vii
Glossary	viii
Acknowledgements	xi
Abstract	xii
Résumé	xiii
0 Introduction	1
0.1 Objectives	2
0.2 Background Information and Motivation for Research	3
0.3 General methodology	8
0.4 Scope and limitations	9
0.4.1 TM systems	9
0.4.2 Texts	10
0.4.3 Languages	10
0.5 Outline	10
1 Introduction to Translation Memories and Fuzzy Matching	12
1.1 Translation memories: From early beginnings to current research	12
1.1.1 Prior research on fuzzy match specifically	15
1.2 How Do TM Systems Work?	16
1.2.1 Information storage in a TM	17
1.2.1.1 Segmentation	17
1.2.1.2 Alignment	18
1.2.1.3 Translation Units	20
1.2.2 Information retrieval from a TM	20
1.2.2.1 Exact Matches	21
1.2.2.2 Full Matches	22
1.2.2.3 Sub-segment Matches	23
1.2.2.4 Term Matches	23
1.2.2.5 Fuzzy Matches	24
1.2.2.5.1 Fuzzy set theory	25
1.2.2.5.2 Measuring similarity: a challenging task	26
1.2.2.5.2.1 Shared words	28
1.2.2.5.2.2 Word order	29
1.2.2.5.2.3 Stop lists	29
1.2.2.5.2.4 Morphological processing	30
1.2.2.5.2.5 Edit distance	30
1.2.2.5.3 Ensuring relevance: balancing precision and recall	31
1.3 Fuzzy Matching and the Translator	33
1.3.1 Selecting a TM	34
1.3.2 Using a TM effectively	35
1.4 Concluding remarks	38

2 Research Questions	39
2.1 The System Factor	39
2.2 The Textual Characteristics Factor	40
2.2.1 Texts containing internal repetitions	42
2.2.2 Recycled texts	42
2.2.3 Revisions	42
2.2.4 Updates	43
2.3 The Language Factor	44
2.4 The Degree of Fuzziness Factor	47
2.5 Concluding remarks	49
3 The System Factor	50
3.1 TM Tool Selection	51
3.2 TM Database Design and Construction	52
3.2.1 TM Database 1: Appeal Decisions	53
3.2.2 TM Database 2: Translation Job Advertisements	55
3.3 Source Text Selection	56
3.4 Testing Procedure	57
3.5 Data Analysis	58
3.5.1 Appeal Decisions TM Database	58
3.5.1.1 Quantitative Analysis	59
3.5.1.1.1 Lack of Fuzzy Matches in Fusion	59
3.5.1.1.2 Disparity in Scores	60
3.5.1.1.3 Disparity with Regard to Multiple Matches	61
3.5.1.2 Qualitative Analysis	62
3.5.1.2.1 Type of Match and Noise Level	62
3.5.1.2.2 Different scores for same segments	64
3.5.2 Job Advertisement TM Database	64
3.5.2.1 Quantitative Analysis	65
3.5.2.1.1 Lack of Fuzzy Matches in Fusion	65
3.5.2.1.2 Disparity in Scores	66
3.5.2.1.3 Disparity with Regard to Multiple Matches	67
3.5.2.2 Qualitative Analysis	67
3.5.2.2.1 Type of Match	67
3.5.2.2.2 Different scores for same segments	68
3.6 Conclusions	68
4 The Textual Characteristics Factor	73
4.1 Text Selection	73
4.1.1 Recycled Texts	75
4.1.2 Revisions	75
4.2 TM System Selection	77
4.3 Testing Procedure	77
4.4 Test Results and Data Analysis	77

4.4.1 Number of Matches Retrieved	79
4.4.2 Average Score of Fuzzy Matches Retrieved	79
4.4.3 Number of High Quality and Low Quality Matches Retrieved	79
4.5 Conclusions	80
5 The Language Factor	83
5.1 Language Selection	83
5.2 CAT Tool Selection	84
5.3 TM Database and Source Text Selection	84
5.4 Testing Procedure	86
5.5 Data Analysis	86
5.5.1 Quantitative Analysis	86
5.5.2 Qualitative Analysis	89
5.6 Conclusions	93
6 The Degree of Fuzziness Factor	97
6.1 Participant Selection	97
6.2 CAT Tool Selection	98
6.3 TM Database Selection	98
6.4 Testing Procedure	99
6.5 Data Analysis	101
6.5.1 Analysis according to participants	101
6.5.2 Analysis according to segments	103
6.5.3 Analysis of participants' comments	104
6.6 Conclusions	105
7 Conclusion	108
7.1 Overall Assessment	108
7.1.1 Chapter 3 – The System Factor	109
7.1.2 Chapter 4 – The Textual Characteristics Factor	109
7.1.3 Chapter 5 – The Language Factor	110
7.1.4 Chapter 6 – The Degree of Fuzziness Factor	111
7.1.5 Challenges Associated with the Development of	112
Guidelines for Translators to Help Them Set the Fuzzy Match	
Threshold	
7.2 Suggestions for Future Research	113
7.2.1 A Holistic Analysis of the Factors	113
7.2.2 Scaling Up the Investigations	113
7.2.2.1 Number of TM Systems Tested	114
7.2.3 Number Categories	114
7.2.4 Number of Languages	114
7.2.5 Larger Participant Pool	114
7.2.6 Size of TM Database	115
7.3 Concluding Remarks	115
References	116
Appendix 3.1 Job Advertisements Fuzzy Matching Results	120

Appendix 3.2 CIC Immigration Appeals Fuzzy Matching Results	127
Appendix 4.1 Text Category Fuzzy Matching Results	136
Appendix 5.1 Language Factor Fuzzy Matching Results	146
Appendix 6.1 Test and Instructions Given to Participants	160

List of Tables

Chapter 1	
1.1 Non-segmented and segmented text extract.	22
1.2 Differences in segmentation that may cause difficulties for automatic alignment	23
1.3 Example of translation units	24
1.4 Example of an Exact Match	25
1.5 Variation that Will Not Result in an Exact Match	26
1.6 Example of Full Match	26
1.7 Example of Full Match	27
1.8 Example of Term Matches	28
1.9 Example of similarity determined by shared words	29
1.10 Example of Stop Lists in English and French	33
1.11 Example of a Fuzzy Match retrieved using a High Sensitivity Threshold	40
1.12 Example of a Fuzzy Match retrieved using a Low Sensitivity Threshold	40
1.13 Multiple fuzzy matches displayed ranging from best to worst according to percentage	41
1.14 An example of how semantic similarity is not reflected in superficial similarity	42
 Chapter 3	
3.1 Comparative Analysis of TM Tools – Immigration Appeals TM Database	62
3.2 Instances where the same segment was retrieved by both systems – appeal decisions	65
3.3 Examples of fuzzy matches generated by Trados that were largely the result of a single sub-segment string match	67
3.4 Comparative Analysis of TM Tools – Job Advertisements TM Database	69
3.5 Instances in which the same segment was retrieved by both TM systems	70
3.6 Examples of fuzzy matches generated by Trados that were largely the result of a single sub-segment string match	72
 Chapter 4	
4.1 Results for “Recycled Texts” (appeal decisions)	82
4.2 Results for “Revisions” (APUO collective agreement)	82
 Chapter 5	
5.1 Number of fuzzy matches produced for each segment in each language direction	91
5.2 Fuzzy match percentage values per segment	92
5.3 Higher English-to-French percentage scores for the same segment	96
5.4 No fuzzy match retrieved from English into French	97
 Chapter 6	
6.1 Results of Multiple Fuzzy Match Experiment	106
6.2 Number of Participants Who Did Not Choose the Highest % Match	107

List of Figures

Chapter 1	
1.1 Misalignment due to one sentence translated by two sentences	23
1.2 Example of a Fuzzy Match	29
1.3 Possible search results from a TM database	36
 Chapter 3	
3.1 Comparative Analysis of TM Tools – Immigration Appeals TM Database	62
3.2 Instances where the same segment was retrieved by both systems – appeal decisions	65
3.3 Examples of fuzzy matches generated by Trados that were largely the result of a single sub-segment string match	67
3.4 Comparative Analysis of TM Tools – Job Advertisements TM Database	69
3.5 Instances in which the same segment was retrieved by both TM systems	70
3.6 Examples of fuzzy matches generated by Trados that were largely the result of a single sub-segment string match	72
 Chapter 6	
6.1 Example of the question format given to translators	104

Glossary

alignment:	The process of building up a translation memory database from previously translated documents by connecting corresponding source and target segments.
bitext:	See parallel corpus.
CAT:	See computer-aided translation.
Computer-Aided Translation:	a process by which humans combine their translation skills with computerized tools to help them translate. Also referred to as <i>CAT</i> .
exact match:	A segment retrieved by a translation memory system that is a perfect (100%) identical match to the corresponding new source language segment. See full match.
expression match:	A term used in the Fusion Translate™ translation memory system to represent a sub-segment match. (See sub-segment match).
full match:	A segment retrieved by a translation memory system that differs from a stored translation unit only in terms of interchangeable information. (e.g numbers, dates, times, currencies, etc.) See exact match.
fuzzy match:	A segment retrieved by a translation memory system that is similar, but not identical to the new source text segment being translated. The fuzzy matching value percentage can fall anywhere between 1% and 99%. Also referred to as <i>partial match</i> .
external repetition:	Segments of text from a translation memory system that are repeated in several different documents.
Internal repetition:	Segments of text from a translation memory system that are repeated within the same document.
leveraging:	See recycling.
localization:	A process through which products or services are customised or somehow adapted linguistically,

	culturally and technically to a particular country or region.
noise:	The information retrieved by a translation memory system that is unhelpful. See also silence.
partial match:	See fuzzy match.
parallel corpus:	A corpus of source language texts aligned with their target language translations. Also referred to as <i>bitext</i> .
recycling:	Reusing portions of previously translated segments in a new target text with the help of a translation memory. Also referred to as <i>leveraging</i> .
revision:	A changed version of a previous source language text. Translation memory systems usually work well with this type of text.
silence:	The information contained in a translation memory database but not retrieved by a translation memory system that could be potentially helpful. See also noise.
sub-segment match:	A match where small portions of contiguous text in the new source text to be translated match up with portions of text in a translation memory database. Also referred to as an <i>expression match</i> in Fusion Translate™
TM:	See translation memory.
translation memory:	A database of previously-translated target language texts and their corresponding source language translations that translators can recycle with the help of a TM system when translating a new text. See also parallel corpus. Also referred to as <i>TM</i> .
translation memory system:	A type of CAT software that translators utilise to recycle previous translations through the creation of a linguistic database, e.g. Fusion Translate® or Trados®.
translation unit:	A segment of source language information aligned with its target language translation and stored as a pair in a translation memory database.

update:

A change in the source text that is made whilst the translation is in progress.

Acknowledgements

First and foremost, my sincere thanks and utmost appreciation go to my thesis advisor, Dr. Lynne Bowker. Dr. Bowker always had time for my questions and concerns, especially at times when both of our lives made it quite difficult to make writing, reading, and revising a thesis a priority. Her understanding and expertise in both translation studies and outside university life made it possible for me to continue and apply myself at times when I was worried about experiment results. Please accept my heartfelt thanks for your dedication to my project.

Another very important person in my life that I would like to thank is my partner David Lee Chung. Thank you for your encouragement and support when I needed it the most and for keeping my stomach filled with delicious meals and pastries during long nights at the keyboard. Without your love and support this thesis would never have come to fruition. *Mo kontan twa!*

I would also like to thank the translator-participants who volunteered their time and know-how: Marie Alnwick, Cathryn Arnold, Hazel Curties, Melissa Ehgoetz, Jean-Frederic Hübsch, Francie Gow, Elizabeth Marshman, Julian Wallis, and Rie Yamagishi. Without their participation, this thesis project would not have been possible.

Abstract

In today's global marketplace, translators are responding to the current demand to produce fast and high-quality translations by using electronic tools to help them do their jobs, and one of the most promising tools that translators have at their disposal is translation memory (TM) software, a veritable database of previously translated material. Fuzzy matching – where the TM system identifies a portion of the text that is similar to but not exactly the same as one stored in the database – has become an integral feature of TM software; yet using this feature effectively remains a mystery to most translators. This is largely because translators have not been presented with any type of guidelines with regard to helping them identify an ideal setting for the fuzzy match value.

The objective of this thesis is to provide translators with a better understanding of TM systems by exploring fuzzy matching in detail, and particularly by investigating factors such as the TM system selected, the category of text being processed, the working languages involved, and the degree of fuzziness of the match. Accordingly, a series of experiments have been designed and carried out to determine the influence that these four factors might have on the ideal fuzzy match setting. The results of these experiments show that these factors should indeed be taken into account when translators are selecting the fuzzy match value to be used with a TM system.

Résumé

Dans le marché mondial d'aujourd'hui, les traducteurs répondent aux demandes actuelles de produire des traductions rapides et de très bonne qualité au moyen des outils électroniques pour les aider à compléter leur tâche. L'un des outils les plus prometteurs dont disposent les traducteurs est un mémoire de traduction, une véritable base de données de textes déjà traduits. Les appariements flous, une partie du texte identifiée par un mémoire de traduction qui est semblable au texte mémorisé dans la base de données du mémoire de traduction, sont désormais une partie intégrante des mémoires de traduction. En revanche, l'usage de cette fonction demeure effectivement un mystère pour la plupart des traducteurs car jusqu'à maintenant, les professionnels langagiers n'ont jamais reçu des directives quelconques pour les aider à trouver un réglage idéal du seuil des appariements flous.

L'objectif de la présente thèse est donc de permettre aux professionnels langagiers de mieux comprendre les mémoires de traduction en explorant en détail les appariements flous, en particulier les facteurs qui peuvent les influencer, tels que le système choisi, la catégorie du texte saisi, la langue ainsi que le degré du flou de l'appariement. Par conséquent, une série d'expériences a été développée et mise à l'essai pour déterminer l'influence que ces quatre facteurs pourraient avoir sur un réglage idéal des appariements flous. Les résultats des expériences montrent que ces facteurs devraient être pris en considération lorsque les professionnels langagiers sélectionnent le degré d'appariement flou d'un mémoire de traduction.

0 Introduction

In recent years, the demand for documentation to be translated into various languages has increased exponentially, and translators today are using electronic tools more than ever to help them do their jobs. In fact, the Canadian Translation Industry Sectoral Committee published a survey in 1999 entitled *Survey of the Canadian Translation Industry* and reported that the area of the translation industry that will experience the most growth – almost 50 per cent per year – is the sector involving technological aids for translators. (1999: 39)

Translators today have a veritable electronic toolbox at their disposal to help them translate faster and more efficiently, and tools such as the internet, electronic dictionaries, and corpus analysis tools have the potential to help translators do their work at a faster pace. In earlier days, translators feared the obsolescence of their livelihood with the development of software that translates documents with little to no human intervention, such as automatic translation systems. However, Schäler (2001:1) claims that such fears can be put to rest since, except for in a few special circumstances¹, “machine translation can safely be declared obsolete and dead.”

¹ Automatic translation systems are used in a few highly specialized and restricted domains (e.g. the Météo system, which is limited to translating weather forecasts from English into French), and they can also be used to produce very rough drafts which may be helpful "for information only" purposes; however, for the vast majority of translation needs, the quality of machine translation output is not sufficiently high unless it is revised by a human translator. Therefore, other types of tools, which aim to help rather than to replace translators, are currently much more popular.

According to Somers (2003b:31), “One of the most significant computer-based aids for translators is the now widely used **translation memory**.” Translation memory systems (abbreviated henceforth as TMs) help translators by using already-translated documents stored in a database compiled by the translator, and comparing the new source text at hand against the previously translated texts in the database. In some instances, a given segment of the source text may have been translated before. When the TM finds the precise segment of the source text already in the database, this is known as an **exact match**. Clearly, not all matches will be exact since not every text is precisely the same. More frequently, when there are similar matches in the database, the TM suggests a segment that differs only slightly from the original. This is referred to as a **partial or fuzzy match**. The degree to which the fuzzy match differs from the original depends on the minimum match value used by the TM – a value that can be set and changed by the system user.

As noted by Somers (2003b:31), TM systems were not commonly available until barely a decade ago. Since translators have been using TM systems for a relatively short time, there are many aspects of their use and impact that are not yet well understood. One of these is the ideal minimum match value that is used in fuzzy matching. With the research in this thesis, I hope to take some steps towards increasing our knowledge in this regard.

0.1 Objectives

The main objective of this thesis is to gain a deeper understanding of one important yet poorly understood feature of TMs – fuzzy matching. This objective will be

achieved by conducting a number of investigations. Firstly, the thesis aims to investigate three factors that may influence the ideal minimum match value for fuzzy matching: the “System Factor” (i.e., which TM system is being used), the “Text Category Factor” (i.e., which type of text is being translated), and the “Language Factor” (i.e., which working languages are involved). An additional study will focus on the “Degree of Fuzziness Factor,” investigating whether or not the fuzzy match with the highest score is always the most helpful match for a translator. Based on the findings generated by these investigations, I will attempt to propose some preliminary guidelines that could help translators to determine the ideal minimum match value to be used in various circumstances.

0.2 Background Information and Motivation for Research

Language professionals have felt the recent boom in technology perhaps more so than any other profession, and the improvements in technology and in translation technology specifically have permanently changed the way translators go about completing their tasks. This is apparent as more and more companies expand their services and sell their products on a global scale, increasing the need for translation, helping to fuel the globalization fire.

Globalization has indeed played an important role in an increase of the use of TMs, and represents another reason why further development and improvement of TMs is necessary. Numerous scholars (e.g. Cronin 2003; Zervaki 2002) have reported how this trend directly affects the field of translation. For instance, large corporations (in particular software companies) increasingly sell their products on a

worldwide scale, participating in a phenomenon known as simultaneous shipment, or *simship*. Simship refers to the release of international, localized versions of software at the same time, or within a short period of time of the original release. (Bowker, 2002b: 198) Simship requires companies to have all language versions of their products including product documentation ready for distribution simultaneously. For example, when IBM launches a new version of a given product, the user manual, technical support material, warranty information, and any other accompanying documentation must be ready to be sent to customers all over the world simultaneously, in all language versions. The idea is to offer these new, updated versions in a variety of languages at the same time. With business practices such as simship, the pressure is on translators now more than ever to stay competitive in a global marketplace by producing their translations at a very high speed, all the while maintaining high quality standards.

The development of the World Wide Web has also increased demands for translation literally around the world. In the early days of the internet, most of the websites and information found online were typically in English, but today an increasing number of companies are realising that their customers are much more liable to make purchases if the information on their website is available in the target language. As observed by Allen (2003: 299), "Large corporations and small-to medium-sized companies alike have been undergoing the expansion of their business to the four corners of the earth. It is no longer possible to rely neither on local business, or to base one's commercial expectations on a single language as the sole medium of communication."

In fact, the International Data Corporation reported that in 2001 over half of the world's 147 million internet users were non-English speakers. It was estimated that by 2005 this figure would have risen to two-thirds. (Cronin, 2003: 14) As the sheer volume of information grows, so do the demands for translation. Allied Business Intelligence reported in 2002 that the "consolidation of translation firms that are occurring through mergers and acquisitions will only make for larger companies that can offer a broader service to fit their customers' needs," all of which is expanding the translation market² and further fuelling the demand for quick turnaround times on translation.

Many translators and employers have realised that TM systems have the potential to help those who translate to continue producing high-quality translations, but at a faster pace as compared to traditional methods (i.e. limited use of electronic tools). However, when translators and/or their employers decide to utilise TM systems, many of them, ranging from novices to experts, are unsure of how to deal with the fuzzy match setting. Most systems have a default setting at around 70% or 75%, and many users simply leave the setting at this default value at all times, simply assuming that this is the optimal setting. Other users may not even realise that they have the option of changing the match value. In those cases where translators do decide to adjust the setting, they may waste a significant amount of time – and ultimately, money – using trial and error methods to adjust the fuzzy matching

² Excerpt from "Language Translation, Localization, and Globalization: World Market Forecasts, Industry Drivers, and eSolutions." ©2002 Allied Business Intelligence Inc

value of their TM system in hopes of finding an ideal setting. This could lead to frustration and could end up with the translator eventually deciding not to use the fuzzy matching option at all. As described by Schäler (2001), working with fuzzy matches “can be a highly complex operation and, in fact, so cumbersome that translators often opt out of the fuzzy match proposal operation by setting the percentage threshold of the fuzzy match component so high that high percentage matches which could contain matching phrases are hidden away from them.”

Most TM systems allow users to adjust the fuzzy matching sensitivity value typically anywhere between 30% and 99%. However, if the wrong setting is employed, fuzzy match results may be either insufficient or overwhelming, problems referred to as *silence* and *noise*³. Sadly, because the fuzzy matching value is often set inappropriately, many users of TM systems waste significant amounts of time, energy, and potential revenue either re-translating portions of segments that are already in the TM database, or retrieving and examining impertinent matches.

Why are translators left to flounder when it comes to making this important decision about the fuzzy match setting? The main reason is because, up until now, there has been no serious empirical research conducted into fuzzy matching values, nor has anyone attempted to establish fuzzy matching guidelines for translators using TM systems. The literature reveals that leading experts on TMs do not make any recommendations as to where to set the fuzzy matching value. Rather, they simply note the fact that the value can be changed by the user:

³ This will be discussed in more detail in section 1.2.2.5.3.

- « L'utilisateur peut parfois régler le pourcentage de divergences ». (L'Homme 1999: 216)
- "user-defined minimum match value." (Austermühl 2001: 137)
- "Most TM systems with fuzzy matching capabilities allow the user to set a threshold where similar segments are retrieved as fuzzy matches." (Esselink 2000: 36)

Where scholars have offered advice on which match value works best, their recommendations vary considerably. For instance, O'Brien says, "From our experience, we have seen that even a 40% match can present useful translation information to the translator." (O'Brien, 1998: 117) However, Yunker claims that "Generally, when new text is processed, fuzzy match parameters are set, usually at 80% or higher, so that only the most likely matches are displayed. Anything below that margin is left for manual translation." (Yunker, 2003: 224) Yet other experts claim that the ideal setting is somewhere between the two extremes, such as between 60% and 70%. (Bowker, 2002a: 100)

Unfortunately, none of these authors qualify their suggestions with regard to what factors might influence a translator's decision to choose a setting of 40% as opposed to one of 80%. Instead of containing helpful guidelines to assist translators in choosing a useful minimum match value, the literature seems to consist primarily of anecdotal experiences that are not based on any empirical research. Cronin (2003: 113) draws attention to the lack of empirical research in translation technology, noting that there is "a dispiriting tendency to substitute anecdote for analysis when certain translators are discussing the impact of computer-assisted translation (CAT) or machine translation (MT)." With this thesis, I

will attempt to take research about TMs, and about fuzzy matching in particular, beyond the realm of anecdotal reporting by conducting an empirical investigation into some of the factors that can influence the ideal match setting. Moreover, based on the empirical evidence gathered, I intend to establish a set of preliminary guidelines that translators can use to help them choose a useful minimum match value for fuzzy matching in TMs.

0.3 General methodology

The overall methodological approach used to conduct the research presented in this thesis is divided into the following sections:

1. Present a literature survey to determine what researchers have already learned about fuzzy matching in the context of TM systems.
2. Design a series of experiments to investigate some of the factors that may affect the ideal minimum match value.
3. Conduct preparatory work, such as building TM databases and generating fuzzy matches.
4. Carry out small scale experiments.
5. Analyze the data generated from the experiments.
6. Draw initial conclusions and write up preliminary guidelines.
7. Assess the overall project, suggest possible modifications to the experimental design, and propose possible directions for future research.

0.4 Scope and limitations

As noted in section 0.1, the aim of this thesis is to investigate the extent to which factors such as the choice of TM system, the nature of the text being translated, the working languages, or the degree of fuzziness of the matches, have an influence on the ideal fuzzy match threshold. However, owing to practical constraints, such as limitations of time and resources, it is necessary to put some limits on this research project.

0.4.1 TM systems

There are currently a considerable number of TM systems available on the market, and most if not all of them employ fuzzy matching techniques⁴. However, it was not feasible to test every single commercially-available TM system. I therefore chose two TM tools, Trados^{TM5}, and Fusion Translate^{TM6}. Both were selected since copies of the software are readily available at the University of Ottawa's Writing Centre computer laboratory. In addition, I had some previous exposure to these two tools through course work and practical work experience, which meant that I had a lower learning curve when it came time to do practical experimentation with these tools. Finally, both systems are relatively well-known in the industry, though Trados is certainly the better known of the two.

⁴ For example, some of the most popular systems include Trados Translator's Workbench, Fusion Translate, Déjà Vu, STAR Transit, SDLX, WordFast, MultiTrans and LogiTrans.

⁵ For more information on SDL International, developers of Trados, see: <http://www.sdl.com/products-home/products-topics/products-translation-memory.htm> [Consulted July 12, 2006]

⁶ For more information on Orca Development Corporation, developers of Fusion Translate, see: <http://www.orcadev.com/> [Consulted July 12, 2006].

0.4.2 Texts

A further project limitation centred on the number and nature of texts used for the experiments. Three TM databases were compiled as part of this research: 1) a database of appeal decisions for nationals seeking permanent residence (landed immigrant status) in Canada type, 2) job advertisements for language professionals, and 3) the collective agreement of the Association of Professors of the University of Ottawa. A more detailed discussion of the reasons for choosing these texts for compiling the TM databases can be found in sections 3.2.1, 3.2.2 and 4.1.2 respectively.

0.4.3 Languages

A final limitation on this thesis project involves language. Even though TMs can be utilised with virtually any language combination, the languages used in this thesis and in the corresponding TM database were limited to English and French, which are my primary working languages, as well as the official languages of the University of Ottawa and of Canada.

0.5 Outline

In addition to this introduction, this thesis contains seven chapters. Chapter 1 presents a brief history of TM systems, an overview of how they work, a detailed explanation of fuzzy matching, and a discussion of why translators need to be concerned with fuzzy matching. Chapter 2 provides a detailed outline of the research questions addressed in this thesis. This discussion sets out the

groundwork for the following four chapters (3, 4, 5 and 6), which contain the descriptions and results of the empirical investigations that were conducted. Chapter 3 presents the experiment concerning the “System Factor” (i.e., the effect of a given TM system on fuzzy matching). Chapter 4 focuses on an investigation of the “Text Category Factor” (i.e., the impact of textual characteristics on fuzzy matching). In Chapter 5, the “Language Factor” (i.e., the effect of the working languages on fuzzy matching) is explored. The subject under investigation in Chapter 6 is the “Degree of Fuzziness Factor” (i.e., whether the fuzzy match with the highest score is always the most useful). Finally, Chapter 7 contains an overall assessment of this research project, some general closing remarks and some suggestions for future research. The thesis is rounded out by a list of references, a glossary, and a series of appendices containing information such as the detailed data generated from the experiments.

1 Introduction to Translation Memories and Fuzzy Matching

1.1 Translation memories: From early beginnings to current research

A translation memory (TM) system can be defined as a linguistic database of source texts aligned with their corresponding target language translations (Bowker, 2002a: 92). Melby (1995: 187) comments that the idea of TM systems has been around for some time, dating back to the 1970s, but early prototypes were not implemented until the 1980s, and it is only since the late 1990s that TMs have become fairly popular among translators.

The initial idea that translators could recycle previous translations in future texts is generally attributed to Martin Kay who in 1980 wrote an influential paper entitled “The Proper Place of Men and Machines in Language Translation.”⁷ Kay (1980[1997]: 19) proposed the idea for software in which

...the translator might start by issuing a command causing the system to display anything in the store that might be relevant to [the text to be translated] ... Before going on, he can examine past and future fragments of text that contain similar material.

Not long after, European Union translator Peter Arthern hypothesized that translators could benefit from electronic access to an aligned corpus of source and

⁷ This paper first appeared as a Xerox PARC Working Paper in 1980, but it has since been reprinted in the journal *Machine Translation*, vol. 12 (1/2), pp. 3-23, 1997.

target texts. What Arthern proposed is more or less a description of what we now know as TMs:

It must in fact be possible to produce a programme which would enable the word processor to “remember” whether any part of a new text typed into it had already been translated, and to fetch this part, together with the translation which had already been translated, ... Any new text would be typed into a word processing station, and as it was being typed, the system would check this text against the earlier texts stored in its memory, together with its translation. ... In effect, we should be operating an electronic “cut and stick” process which would, according to my calculations, save at least 15 per cent of the time which translators now employ in effectively producing translations. (Arthern, 1981: 318)

Melby (1995: 225) further suggests that the first implementation of a prototype TM system originated under his direction at Brigham Young University (BYU) in the United States in the 1980s. Since the late 1990s, we have witnessed a veritable explosion of commercial systems on the market, including TRADOS Translator’s Workbench, MultiTrans, Déjà Vu, WordFast, STAR TransIT, SDLX, and Fusion Translate, among others.

Because these tools have become so incredibly popular, researchers are finally beginning to show an interest in learning more about them. According to Schäler (2001: 1), when the idea of a TM was first proposed, “computational linguists often chose to simply ignore TM technology – considering it as some type of sophisticated search and replace engine, not a subject for serious research efforts.” Now, however, it is clear that TM technology is here to stay, and it can no longer be ignored by

researchers since all anecdotal evidence indicates that this technology is having a significant impact on both the translation process and product.

One of the main areas under investigation deals with the question of productivity. Researchers such as Webb (1998), Bowker (2005) and Circé (2005) have compared the use of a TM system against unaided human translation and demonstrated that, when used in the right circumstances, TMs can lead to significant productivity gains. Bowker (2005) also investigates the question of quality, examining whether the use of a TM system adversely impacts the quality of the target text. She concludes that if used carelessly, a TM can indeed lead a translator to produce a lower-quality text, warning that it is important to strike a balance between productivity and quality since one can come at the expense of the other. Similar quality-related findings are reported by Bédard (2000) and Mogensen (2000). They warn translators against simply adopting all the proposals offered by a TM system in a slavish fashion and advise translators to remember that they are producing a text and not merely a string of sentences.

Other researchers have studied the current crop of TM tools, noting their strengths and weaknesses as compared to other types of translation tools, such as concordancers (Bowker and Barlow 2004) and machine translation systems (Circé 2005). Yet others have examined the shortcomings of existing TM systems, and suggested ways to improve the next generation of tools. For instance, Macklovitch and Russell (2000) recommend improvements in morphological parsing and whole text preservation, while Bowker (2002b) suggests that thesauri could be

incorporated into TM systems to facilitate the identification of synonyms. Furthermore, as the number of TMs on the market continues to increase, several comparative evaluations of a series of different TM tools have been conducted (e.g. Benis 1999, 2000; Zeffass 2002b) and a number of evaluation metrics have been drafted (e.g. Rico 2000; Höge 2002; Zeffass 2002a; Gow 2003) in an attempt to provide some guidance for helping translators to pick the TM system that best meets their needs.

For the most part, the research conducted on TMs to date has taken a very holistic view of these products. Relatively little attention has been paid to the specific feature known as fuzzy matching, which is a feature that allows a TM database to identify parts of previously translated texts that resemble, but are not identical to, a new source text that is to be translated. When fuzzy matching is mentioned in the literature, it is usually only as a passing reference, or to indicate that it is a complex issue that will not be discussed in detail.

1.1.1 Prior Research on Fuzzy Match Specifically

To the best of my knowledge, the only research that focuses specifically on fuzzy matching is reported in a paper by de Vries (2002) that investigates the issue of payment and TMs. In this paper, de Vries explores the issue of translators receiving reduced payment for translating segments for which a fuzzy match is retrieved from the database, pointing out a number of pitfalls associated with such a policy (e.g., texts may not undergo a “high-level” review or sanity check). This is certainly an

important issue, but de Vries's paper does not provide an in-depth exploration of the fuzzy matching process per se. As a result of this lack of attention, fuzzy matching remains an area that is not well understood. Consider, for instance, the following comments made by Macklovitch and Russell (2000:139):

How exactly do these fuzzy matching algorithms work? It is difficult to say with certainty because TM vendors, although they do illustrate the concept in their promotional literature and demos, do not generally provide a formal definition of the similarity coefficient that users may specify in order to constrain the search for approximate matches. Hence it is not all that obvious how the results of a 70% match will differ, say, from a 74% match or an 81% match. [...] Combining several distinct and incomparable factors into a single numerical measure may appear to simplify things for the user, but as a consequence users are left with a vague and ill-defined comprehension of a parameter that is central to the system.

One of the aims of this thesis is to go some way towards filling this gap by exploring the impact that individual TM systems, textual characteristics, and language can have on a translator's choice of a fuzzy match value. However, before embarking on a detailed exploration of fuzzy matching, it is first necessary to introduce some key concepts and definitions regarding TMs in general. In the following sections, an overall explanation of how TM systems work will be provided. This will be followed by a more in-depth look at fuzzy matching in particular, which includes a brief investigation of fuzzy set theory and information retrieval.

1.2 How Do TM systems Work?

TM systems have two basic functions. On the one hand, they store previously translated texts, and on the other, they retrieve relevant information from these texts. Therefore, the key to understanding how a TM works is to understand how it both stores and retrieves information.

1.2.1 Information storage in a TM

Information that is stored in a TM is organised in a very precise way. Important concepts that must be understood in this regard include segmentation, alignment, and translation units.

1.2.1.1 Segmentation

TM systems process pieces of information known as **segments**, which are most often sentences, but which could also be sentence-like units such as headings, subtitles, or captions. Deciding what forms a segment may seem like a simple task, but that task is complicated by grammatical elements that may “trick” the TM into breaking a segment in the wrong place. For instance, abbreviated forms such as *Dr.*, *Mrs.*, *U.A.E.*, or decimal numbers or an ellipsis (...), etc. might do this to a TM, since periods do not always function as sentence terminators.

Although the results may not be perfect and may require human intervention to correct errors, a TM system nevertheless begins by attempting to divide a text into segments, as illustrated in Table 1.1. This process is carried out on both the source and target text individually.

Excerpt from text prior to segmentation	L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. Il n'a pas contesté la validité juridique de la décision. Il a obtenu le droit d'établissement en 1977 et a établi sa famille, mais vivait en Angleterre depuis 1978. Il n'était pas solidement enraciné au Canada.	
Same excerpt divided into segments	10	L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés.
	11	Il n'a pas contesté la validité juridique de la décision.
	12	Il a obtenu le droit d'établissement en 1977 et a établi sa famille, mais vivait en Angleterre depuis 1978. Il n'était pas solidement enraciné au Canada.

Table 1.1 – Non-segmented and segmented text extract.

1.2.1.2 Alignment

Once the source and target texts have been segmented individually, the next step in the process is **alignment**, which refers to the task of matching up the corresponding source and target segments. We saw in section 1.2.1.1 that segmentation is not always a straightforward process, and the same applies to alignment.

For example, a translator may take a long source text (ST) sentence and translate it by two shorter sentences in the target text (TT), or two short ST sentences may be combined into a single longer sentence in the TT as illustrated in Table 1.2. Moreover, information may be added (e.g. an explanation of a cultural concept unfamiliar to the TT audience) or omitted (e.g. a cultural reference not relevant in the new TT), or the order of sentences may be changed.

One sentence translated by two sentences	FR: Vous trouverez plus de renseignements sur la résidence dans la brochure Comment devenir résident canadien ainsi que sur le web. EN: Read more about permanent residence in the brochure How to Become a Canadian Citizen. Information is also available on the web.
Two sentences translated by one sentence	FR: Utilisez la calculatrice de la période de résidence. Vous saurez donc si vous répondez aux exigences en matière de résidence en vue d'obtenir la citoyenneté canadienne. EN: To find out if you meet residence requirements to become a Canadian citizen, you can use the residence calculator.

Table 1.2 – Differences in segmentation that may cause difficulties for automatic alignment.

In order for the TM to function effectively, corresponding segments must be correctly aligned. As shown in Figure 1.1, if a text is misaligned, the “match” retrieved by the system will be incorrect and will therefore be of no use to the translator. In fact, the translator may even waste time by consulting this incorrect information. Therefore, human intervention may be necessary to recognise and correct any alignment errors that occur.

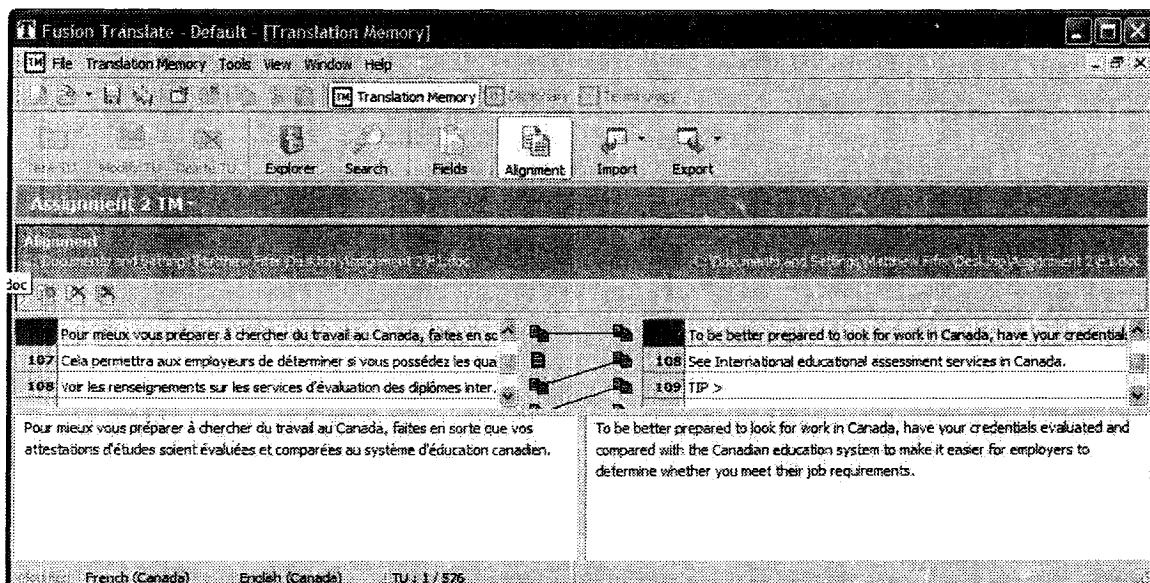


Figure 1.1 – Misalignment due to one sentence translated by two sentences

1.2.1.3 Translation Units

Once the source and target texts have been properly segmented and aligned, the resulting “pairs” of segments are stored in the database in the form of translation units (TUs), as seen in Table 1.3. These TUs are what the system will eventually retrieve and display for the translator to consult whenever a match is found.

Translation unit 1	EN: Before you apply you should make sure that you are eligible for citizenship. FR: Avant de présenter une demande, vous devriez vérifier si vous êtes admissible à la citoyenneté.
Translation unit 2	EN: Please do note that time spent in prison, on parole, or on probation may not be counted towards becoming a citizen. FR: A noter que le temps que vous aurez passé en prison, en liberté conditionnelle ou en probation, le cas échéant, pourrait ne pas être pris en compte aux fins de l’octroi de la citoyenneté canadienne.

Table 1.3 – Example of translation units

1.2.2 Information retrieval from a TM

Now that we understand how information is stored in a TM database, we can move on to a discussion of the information retrieval process. When a translator begins working on a new ST, this text must first be opened in a text editor, such as a word processor, which interacts with the TM system. The TM system then proceeds to divide this new ST into segments, as described in section 1.2.1.1 above. The TM then takes the first segment of this new ST and compares it against all the segments stored in the TM database.

This comparison is done primarily on the basis of what Bowker (2002b:198) calls “superficial character string matching” or what Somers (2003b:38) refers to as

“character-string similarity.” This means that the spelling, punctuation and all other parts of the segments are taken into account when the TM system tries to find matches in the database. When a match is established, the TM retrieves the matching TU from the database and presents it to the translator, who can then choose to accept, modify, or reject this proposal. There are a number of different types of matches that a TM system can retrieve, and these will be discussed below.

1.2.2.1 Exact Matches

An **exact match** (also referred to as a “perfect match”) is exactly what the name suggests – it is a 100% identical character match to the segment to be translated. For a TM to categorise a segment as exact, every aspect of the two segments must be identical; there can be absolutely no difference in spelling, punctuation, formatting (bold, italics, etc.), inflection, or numbers, although some TM systems are starting to become more flexible in this regard. Table 1.4 illustrates an exact match.

<i>Segment from new ST</i>	Find out about applying for permanent residence.
<i>Translation unit from previous translation</i>	EN: Find out about applying for permanent residence. FR: Voyez comment demander la résidence permanente.

Table 1.4 – Example of an Exact Match

However, exact matches are not always easy to obtain for a variety of reasons. Table 1.5 provides some examples of TUs that a TM system will not recognise as an exact match.

Spelling variations	The office will be closed on Labor Day. The office will be closed on Labour Day.
Inflection variations	Bring your passport photo with you. Bring your passport photos with you.
Punctuation variances	Above all, do not forget your passport. Above all, do not forget your passport!
Formatting variances	Do not forget your passport. Do not forget your passport.
Number variations	Fill out form 1.2. Fill out form 1.3.

Table 1.5 - Variation that Will Not Result in an Exact Match

1.2.2.2 Full Matches

As illustrated in some of the examples from Figure 1.6 above, small differences between the new ST segment and the TU that is stored in the database (e.g. *Labor/Labour* or *photo/photos*) may prevent exact matches from being retrieved. Some TM software developers have responded to this problem by providing an additional type of matching called full matching. A **full match** is displayed by a TM when the new source segment is highly similar to a previously stored translation unit, but the difference lies in so-called “variable elements” in the segment such as numbers, dates, times, measurements, or currencies (i.e., elements that do not affect the linguistic translation per se). The example in Table 1.6 below illustrates a full match.

New ST segment	You must send your application for permanent residence by 1-30-06 at 5 P.M.
Stored TU	EN: You must send your application for permanent residence by 12/30/05 at 17:00. FR: Vous devez envoyer votre demande de résidence permanente avant le 30 décembre 2005 à 17h.

Table 1.6 - Example of Full Match

1.2.2.3 Sub-segment Matches

As pointed out by Schäler (2001: 2), among others, repetition in language is more likely to occur at the sub-sentence level than at the sentence level. This has prompted some TM system developers to offer another type of match known as a **sub-segment match**.

Sub-segment matches occur when small chunks of contiguous text in the new ST match up with small chunks of text in a TM. This way, even if a TM cannot propose a match for an entire segment, the translator might at least get help for translating part of the sentence. Table 1.7 below demonstrates a sub-segment match. The TM system Fusion refers sub-segment matches as “expression matches.”

New ST segment	La demande parrainée de résidence permanente de la partenaire conjugale de l'appelant a été refusée.
Stored TU	EN: The appellant appealed the refusal of the sponsored application for permanent residence of her conjugal partner. FR: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son partenaire conjugal.

Table 1.7 – Example of a Sub-segment (expression) Match

1.2.2.4 Term Matches

In a similar vein, repetition in a text might occur at the level of individual terms, rather than larger chunks. For this reason, most TM systems also work in conjunction with an associated term base. This bilingual or multilingual term base must be compiled separately by a translator and it can then be consulted along with the TM database. In this way, if no matches are found in the TM database, the

translator might at least be able to find a **term match** in the term base as illustrated in Table 1.8.

New ST segment	In order to be granted <u>citizenship</u> , your <u>application</u> must be first approved by the province.
Term base entry	EN: citizenship FR: citoyenneté (f)
Term base entry	EN: application FR: demande (f)

Table 1.8 - Example of Term Matches

1.2.2.5 Fuzzy Matches

In the preceding sections, we have seen that some forms of matching attempt to retrieve matches for complete or nearly complete segments (i.e., exact and full matches), while others focus on relatively small contiguous pieces of a segment (i.e., sub-segment and term matches). But what about those portions of a text that fall somewhere in between? Segments in the TM database that are somewhat similar to a new ST segment may be very useful to translators, so most TM programs also offer a powerful feature called fuzzy matching.

As illustrated in Figure 1.2, a **fuzzy match** is a partial match. It differs from a full match in that the differences are not restricted to variable elements, and it differs from a sub-segment match in that the points of similarity need not be contiguous but can be spread throughout the segment.

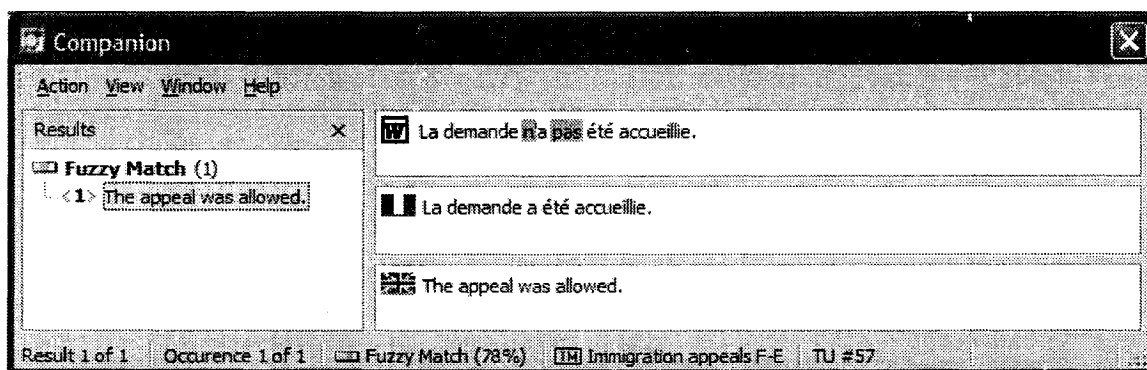


Figure 1.2 – Example of a fuzzy match

A translator can edit a fuzzy match as necessary in order to produce a valid translation for the new ST segment. The parts of the retrieved segment that differ from the new ST segment are usually highlighted to indicate what needs to be modified in the current translation in order to match the corresponding original. Some TM systems use different colours to indicate different types of changes that need to be made to the target text (e.g. additions, omissions, modifications).

While the computational means used to identify exact, full, sub-segment and term matches are fairly straightforward (i.e., the computer engages in more or less strict pattern matching), the identification of a fuzzy match is a more complicated matter. To understand how fuzzy matching works, we need to understand some basic points about fuzzy set theory and information retrieval.

1.2.2.5.1 Fuzzy set theory

Fuzzy matching has its roots in fuzzy set theory, a mathematical concept introduced by University of California (Berkeley) professor Lotfi Zadeh in 1965.

(Ross/Booker/Parkinson, 2002: 4) In layman's terms, fuzzy set theory is used to express knowledge that cannot be expressed in Boolean (e.g., true/false, yes/no) terms. To illustrate this idea, consider the following scenario, adapted from an example given in the Wikipedia entry for "Fuzzy Logic"⁸:

John has a house. Much of the time, John's status within the set of "in the house" is clear: he is either "in the house" or "not in the house." But what happens when John stands in the doorway? He may be considered "partially in the house." Quantifying this partial state produces a fuzzy set membership. For example, with only his little toe outside, we might say John is 99% "in the house" and 1% "not in the house."

As pointed out by Schwarzl (2001: 213), fuzzy set theory deals with vagueness and imprecision. This is why it can be usefully applied in TM systems in the form of fuzzy matching – a type of matching algorithm that looks for vaguely similar, but not identical, segments.

1.2.2.5.2 Measuring similarity: a challenging task

We have seen in section 1.2.2.1 that exact matches operate on Boolean principles where the new ST segment is either exactly the same as the stored segment (i.e., "in the house") or it is not (i.e., "not in the house"). The sub-segment and term matches described in sections 1.2.2.3 and 1.2.2.4 also operate on Boolean principles, but using smaller portions of text. However, in order to get the most out of a TM system, translators also need to be able to identify and retrieve those sentences that are "standing in the doorway," so to speak. Essentially, this means

⁸ http://en.wikipedia.org/wiki/Fuzzy_logic

that TM systems need to have some means for measuring similarity between sentences.

As pointed out by Trujillo (1999: 61), this is actually a very difficult task for humans, let alone computer systems. Ideally, the measure should take into account both semantic and syntactic similarity; however, as noted in section 1.2.2 above, TM systems operate primarily on superficial similarity. They are not intelligent and have no way of understanding the semantic meaning of a sentence. Numerous researchers, such as Macklovitch and Russell (2000), Bowker (2002b) and Somers (2003b), have pointed out this weakness in TM systems, which can be illustrated with an example borrowed from Macklovitch and Russell (2000: 139):

- (1) The wild child is destroying his new toy.
- (2) The wild chief is destroying his new tool.
- (3) The wild children are destroying their new toy.

In this example, sentence (3) is more similar to sentence (1) in terms of semantics and it would be more useful to a translator; however at the level of superficial similarity, it differs from sentence (1) by a total of nine characters. In contrast, sentence (2) is semantically quite different from sentence (1), but at a superficial level, it differs by only four characters.

Faced with this type of challenge, how can similarity be measured by a computer system? A number of different approaches, and combinations of approaches, have

been integrated into various TM systems, including shared words, order of words, stop lists, morphological processing, and edit distance. However, as pointed out by Macklovitch and Russell (2000: 139) and by Somers (2003b: 38), it is unfortunately not possible to know precisely which approaches are employed by which TM systems since the developers of each system consider this to be proprietary information.

1.2.2.5.2.1 Shared words

The most basic method of determining similarity is to count the number of words that the new ST segment has in common with the stored TUs. One problem here is that the length of the sentence may have an effect on the results, as illustrated in the example shown in Table 1.9 borrowed from Trujillo (1999:62).

New ST segment	Move this data to the chart.	
TU 1	You can always move this and other relevant data to the package and display it as a chart.	6 words in common
TU 2	You can move multiple data to the chart.	5 words in common

Table 1.9 – Example of similarity determined by shared words

In this example, TU 2 resembles the new ST segment more closely, but it has fewer words in common than TU 1.

1.2.2.5.2.2 Word order

One possible way to combat the problem described in section 1.2.2.5.2.1 is to add ordering information to the calculation of similarities. For example, the two strings “the number format” and “format the number” would both receive the same score with regard to the number of words they have in common, but if the order in which the words appear is built into the similarity measure, then the latter would be ranked higher than the former if the system were seeking a match for a segment such as “Format the number accordingly.”

1.2.2.5.2.3 Stop lists

Trujillo (1999: 63) notes that since “many of the most frequent words in a language are very poor discriminators of meaning,” one way to improve the similarity measure is to incorporate a **stop list**. According to Bowker (2002a: 153), a stop list is a list of lexical items that are to be temporarily ignored by the computer during certain types of processing, such as the calculation of similarities in fuzzy matching.

Stop lists can be usefully employed in TM systems because the prevalence of function words such as *the*, *of*, *a*, *an*, *in*, and *on* would cause most TM systems to propose matches that had little in common with the new ST at a semantic level. Words typically included in a stop list are articles, conjunctions, prepositions and sometimes adverbs. Table 1.10 illustrates possible examples of stop list items for English and French.

EN stop list	a, an, the, and, or, but, by, from, of, that, to, with
FR stop list	à, au, à la, aux, le, la, l', les, et, ou, mais, par, de, du, d', des, que, un, une, avec

Table 1.10 - Example of Stop Lists in English and French

1.2.2.5.2.4 Morphological processing

Macklovitch and Russell (2000), Bowker (2002b) and Somers (2003b) have all pointed out the fact that morphology is an important factor that should be taken into account when establishing similarity. For example, segments containing “fishes” or “fished” would likely be more helpful to a translator trying to translate the verb “fish” than would a segment containing the word “dish,” even though “fish” and “dish” have a closer superficial similarity (a difference of one character) than do “fish” and “fishes/fished” (a difference of two characters).

1.2.2.5.2.5 Edit distance

Edit distance (sometimes called Levenshtein Distance after the Russian mathematician who, in 1965, discovered the most efficient way to calculate it⁹) refers to the process of measuring similarity by counting the minimum number of insertions, deletions and substitutions required to convert one string of characters into another. For example, the “cost” involved in converting the string “fish” to “dish” is 1 because one substitution (“d” for “f”) is required. Similarly, the cost of converting “fish” to “fishing” is 3 because three additions need to be made. One of

⁹ http://en.wikipedia.org/wiki/Levenshtein_distance

the most common applications of edit distance is in spell-checkers, but of course, it can also be usefully applied to fuzzy matching in TM systems.

1.2.2.5.3 Ensuring relevance: balancing precision and recall

The preceding sections have described a number of possible approaches that a TM system can employ to measure the similarity between a new ST segment and segments that have been previously stored in a TM database. Regardless of the specific approach or combination of approaches used, the goal of any TM system is help a translator identify and retrieve useful information from the TM database. Information retrieval is not unique to TMs. There are many other examples of information retrieval systems that translators are likely to encounter in their work, such as a library catalogue system, an Internet search engine, or a concordancer. All forms of information retrieval face a common challenge: they strive to find all the pertinent information while ignoring the irrelevant material.

In the case of fuzzy matching, for example, four possible groups could result from any given search:

- ♦ (A) Relevant matches that were retrieved by the system
- ♦ (B) Irrelevant matches that were retrieved by the system
- ♦ (C) Relevant matches that the system missed
- ♦ (D) Irrelevant matches that were not retrieved by the system

	Relevant	Irrelevant
Retrieved	A	B
Not Retrieved	C	D

Figure 1.3 – Possible search results from a TM database
(Source: adapted from http://www.diagnosticstrategies.com/info_retrieval.htm)

Clearly, group (A) (i.e., relevant matches retrieved) and group (D) (i.e., irrelevant matches not retrieved) are the only desirable results. The problem presented by group (B) (i.e., irrelevant matches retrieved) is generally described as “noise,” and it causes a translator to waste time by consulting irrelevant material. Meanwhile group (C) (i.e., relevant matches not retrieved) presents a problem typically known as “silence” where there is helpful information contained in the TM database, but it is not presented to the translator. This might also waste time since the translator will have to conduct additional research to find a solution even though one already exists in the TM.

In information retrieval, the problems of noise and silence are directly related to two important concepts: precision and recall. Chu (2003: 198) defines precision as “the ratio between the number of relevant documents retrieved and the total number of

documents retrieved from a system.” Therefore, if a system retrieved four segments from a TM, but only one was useful, then precision is 25%. Meanwhile, recall can be defined as “the ratio between the number of relevant documents retrieved and the total number of relevant documents in a system” (Chu 2003: 198). This means that if one relevant segment is retrieved, but there are two relevant segments in the TM database, then recall is 50%. The difficulty arises because increasing one measure sometimes means decreasing the other. For example, as explained by Trujillo (1999:63), adding ordering information to a search (as described in section 1.2.2.5.2.2 above) might improve precision, but it might simultaneously decrease recall. Therefore, when implementing fuzzy matching algorithms, TM developers must strive to achieve a balance between precision and recall in order to generate the best overall results.

1.3 Fuzzy matching and the Translator

The preceding sections have provided a detailed description of how fuzzy matching algorithms actually work, but issues such as deciding which similarity measures to implement and trying to achieve a balance between precision and recall are the remit of TM system developers. Since translators really have no say in those matters, this might beg the question as to why they should be concerned with fuzzy matching. I submit that there are at least two reasons why translators would benefit from knowing more about fuzzy matching. The first reason is because it could help them to decide which system to select, and the second is because it might help them to use this system more effectively.

1.3.1 Selecting a TM

As mentioned in section 1.1, there are currently a vast number of TM systems available on the market. This has prompted a number of translators, such as Benis (1999, 2000) and Zerfass (2002b), to undertake comparative reviews of many of the current systems with a view to assessing the strengths and weaknesses of each. While these evaluations do acknowledge that fuzzy matching exists in all systems, they do not comment on it in any detail. In fact, Zerfass (2002b:14) states that because the fuzzy matching algorithms used by each system are a little different, it is not really possible to compare this feature of the systems. Instead, these comparative evaluations focus on issues such as user-friendliness of the TM system, what file formats can be processed, how well the systems can perform segmentation and alignment, and how much they cost.

Some researchers, such as Rico (2000) and Höge (2002), have tried to go further by attempting to develop metrics that translators could use to evaluate any of the available systems to see which best meets their needs. However, even here, the degree to which fuzzy matching is truly evaluated is limited. Rico (2000:37), for instance, suggests that it is very important to consider “accuracy” when evaluating TM systems, where she defines accuracy as a measure of “system performance in terms of precision (percentage of valid segments from all those retrieved) and recall (percentage of segments retrieved from all those valid in the TM database).” However, she does not provide any concrete details as to how a translator could

actually go about measuring and evaluating system accuracy with regard to fuzzy matching, nor does she provide any examples. Clearly, more work remains to be done in this area.

1.3.2 Using a TM effectively

Once a TM system has been selected, it is critical for a translator to understand the fuzzy matching feature in order to make the best possible use of the system.¹⁰ When working with a TM system, a translator can adjust the sensitivity of the match threshold. As pointed out by Höge (2002: 171): “If the minimum match value in the system is set to 30%, more matches with presumably a lower quality will be retrieved than if it is set to 60%.” Clearly then, translators have an important decision to make. On the one hand, if they set the match value too high, they may have good precision but poor recall, resulting in “silence.” However, on the other hand, if they set the value too low, they will have high recall but poor precision, resulting in “noise.”

Tables 1.11 and 1.12 show examples of fuzzy matches obtained using high and low sensitivity thresholds.

¹⁰ Note that while a translator may sometimes have the luxury of selecting a system to work with, often the choice is made by the client or the company, in which case, a translator really needs to know how to make the most of whatever system is at his or her disposal.

New source segment	Voici des renseignements inutiles sur la façon de devenir résident canadien.
Stored TM unit	FR: Voici des renseignements utiles sur la façon de devenir résident canadien. EN: Here is useful information about becoming a Canadian resident.

Table 1.11 - Example of a Fuzzy Match retrieved using a High Sensitivity Threshold

New source segment	Voici des renseignements inutiles sur la façon de devenir résident canadien.
Stored TM unit	FR: Pour recevoir des renseignements sur la résidence permanente, demandez auprès du consulat canadien. EN: For information on permanent residence, contact the Canadian Consulate.

Table 1.12 - Example of a Fuzzy Match retrieved using a Low Sensitivity Threshold

Although most TM systems come with a default setting – usually around 70% or 75% – they also allow the user to change this setting. As noted in section 1.2.2.5.1 above, when John is standing with only his little toe outside, he can be considered to be 99% “in the house” and 1% “outside of the house.” The same holds true with fuzzy matches – they can be anywhere from 1% to 99% similar to the new ST segment. Experience has shown that matches that are established with between 1% and 30% similarity are nearly always made on the basis of function words (e.g., *a, an, the, to, of, and*). Since translators do not typically need help translating function words, retrieving such matches is a waste of time. Therefore, most TM systems only allow the match threshold to be set somewhere between 30% and 99%.

Nevertheless, this still leaves the translator with quite a range to work with. How can he or she decide the best match value to use? As noted in section 0.2, the literature provides only anecdotal evidence, and this ranges from advice to set the match threshold at 40% (O'Brien 1998: 117) to a recommended setting of 80% (Yunker 2003: 224), and many points in between. Translators could clearly benefit from having some firmer guidelines to help them make this important choice.

The decision about where to set the match threshold is further complicated by the fact that many TM systems are capable of finding more than one fuzzy match (Brockmann 1999: 9). In such a case, the system will display all matches in an order ranging from best to worst, where “best” is the match that bears the closest superficial resemblance to the new ST segment to be translated, as illustrated in the following example borrowed from Brockmann (1999:9), which has been reproduced in Table 1.13.

Original source text segment	What exactly is a <i>translation memory</i>?
86% fuzzy match retrieved from TM database	What precisely is a <i>translation memory</i>?
72% fuzzy match retrieved from TM database	What is meant by <i>translation memory</i>?

Table 1.13 – Multiple fuzzy matches displayed ranging from best to worst according to percentage (source: Brockmann 1999: 9).

The reason that this could present a problem is because, according to researchers such as Macklovitch and Russell (2000: 138-139), Bowker (2002b: 200) and

Somers (2003b: 39-40), it is possible that the second (or third, etc.) match *might* actually be closer to what the translator needs or would like to insert into the new translation. This could happen because, as mentioned in section 1.2.2.5.2, a TM ranks matches not on the basis of semantic similarity but rather on the basis of the superficial similarity of the segments. For instance, consider the example in Table 1.14, borrowed from Bowker (2002b), which demonstrates how semantic similarity may not be reflected in superficial similarity.

New ST segment to be translated	File the form.
“Best” fuzzy match retrieved by TM (difference of 2 characters)	Fill the dorm.
“2 nd best” fuzzy match retrieved by TM (difference of 7 characters)	He is filing the form.

Table 1.14 – An example of how semantic similarity is not reflected in superficial similarity
(Source: Bowker 2002b)

In this example, the “best” match has the closest superficial similarity to the new ST segment, but the “2nd best” match would actually be more helpful to the translator. Therefore, translators may need to keep this in mind when setting their match value since ironically, a lower percentage match might turn out to be more helpful than a higher one. ¹¹

1.4 Concluding remarks

In this chapter, the concepts of translation memory and fuzzy matching have been introduced and explained. The information presented here will form part of the basis for a number of research questions that will be presented in Chapter 2.

¹¹ This hypothesis will be explored and tested in Chapter 6.

2 Research questions

Keeping in mind the objectives outlined in section 0.1, and based on the knowledge that I have acquired from the review of the literature on TMs and fuzzy matching, as well as on my own practical experience with TM tools, I will now formulate some research questions for investigation. These questions address four different areas, which I refer to as 1) the “System Factor,” 2) the “Text Category Factor,” 3) the “Language Factor,” and 4) the “Degree of Fuzziness Factor.”

2.1 The System Factor

As discussed in section 1.2.2.5.2, numerous authors (e.g., Trujillo 1999: 61; Macklovitch and Russell 2000: 139; Zeffass 2002b: 14; Somers 2003b: 38) have commented on the fact that there are multiple approaches that can be used to measure the similarity of segments (e.g., shared words, word order, stop lists, morphological processing, edit distance). It is also well-known that each TM system uses a different approach or combination of approaches. However, as pointed out by Somers (2003b: 38), the specific details are not made public: “How this score [the fuzzy match score] is arrived at can be quite complex, and is not usually made explicit in commercial systems, for proprietary reasons.”

While people are quick to point out that different systems produce different results, no one has yet, to the best of my knowledge, attempted to undertake any type of systematic study to see if it can be shown that, for example, TM system A

consistently produces a higher fuzzy match score than TM system B for the same set of matches. If a comparative table of the fuzzy match scores of different TM systems could be drawn up, it would be of enormous help to translators when it came time for them to set the fuzzy match value. They would be able to see, for instance, that if they were using TM system B, they would need to set the match value to a lower percentage than they would if using TM system A. This type of information could be very valuable to translators because, as noted by Wallis (2006: 10), translators do not always get to choose the system they work with. Sometimes, a client will insist that a particular system be used, and if the translator is not very familiar with that system, s/he may not realise that in order to achieve results that are comparable to that of a different system, a higher or lower match value may be required on the other system.

Therefore, as part of this research project, I will attempt to comparing the fuzzy match results of two different TM systems with a view to determining if one system systematically outperforms the other.

2.2 The Text Category Factor

Text typology is a very contentious issue in Translation Studies and elsewhere. The idea behind a text typology is that it is possible to classify texts into different categories according to various characteristics. The characteristics that are deemed to be important depend primarily on the goal or purpose of the classification. There is no definitive typology or single “correct” way of classifying texts. Many different typologies have been proposed and used, and they have been based on criteria such

as target audience, function, domain, content, etc. (e.g. Trosborg, 1997; García Izquierdo, 2000). These are all interesting approaches, which are valid in certain circumstances, but they are not highly relevant in the context of TM systems. As explained in section 1.2.2.5.2, TM systems are not intelligent tools. Rather, they operate on the basis of pattern matching, attempting to match the superficial character strings found in the new source text against those found in the texts that are stored in the TM database. For this reason, a typology based on target audience or text function is not very useful because a TM system is not capable of identifying these elements. In the context of this research project, I feel that it would be excessive to create a detailed and full-blown text typology. Because the TM system is looking for superficial matches, what is actually needed is an analysis of surface characteristics that can be found in a text that either do or do not make that text amenable to being processed with the help of a TM system.

Much of the literature on the subject of text suitability for TM processing is relatively vague. For instance, authors such as Austerlühl (2001: 138) and Quah (2006: 193), suggest that scientific, technical or semi-technical, administrative and informative texts are better suited for TM processing, while literary and creative texts should be avoided. Meanwhile, other authors propose simply that TMs are most useful for texts that are “de natures semblables” (L’Homme 1999: 351) or texts that “resemble each other” (Trujillo 1999: 59).

Two authors who give slightly more detailed descriptions of texts that can most usefully be translated with the help of a TM system are Webb (1998: 14) and

Bowker (2002a: 112-114). These authors identify the following four categories of texts as being conducive to TM processing.

2.2.1 Texts containing internal repetitions

These are texts that contain a great deal of repetitive content, such as legal texts, for example. The reason these texts are considered good candidates for a TM is that as soon as a given segment is translated and stored in the TM database, it becomes available for reuse and can even be used to help translate the very next segment of the new source text.

2.2.2 Recycled texts

These are texts that come from the same client and/or the same subject field and/or the same genre. In such cases, the texts are likely to contain terminological or stylistic preferences (from a client) or similar terminology (from the same subject field) or similar structure/format (from a given genre). In other words, there may be passages in the new source text that are similar to passages in other texts (e.g. fixed expressions, terminological phrases) and which can therefore be recycled to some degree.

2.2.3 Revisions

These are amended versions of a previous text. A good example of a revised text is a user manual. Imagine that a manual for a product has been written and translated. At a later date, some new features are added to the product and

corresponding descriptions of these new features are added to the manual. The revised manual has a lot of overlap with the original manual – the only difference is the descriptions of the new features. Now imagine that the translator who translated the original manual had stored it in a TM. Instead of retranslating the entire manual, the translator can reuse the previous translation for the unchanged portion of the text and then concentrate on translating the new additions.

2.2.4 Updates

These are texts that are not in their final form. An update occurs when a client makes changes to a source text while the translation is still in progress. Receiving an update can pose difficulties for a translator, especially if the text is long and the changes have been made throughout the document. However, with a TM system, the translator can run the updated source text through the system, and it will quickly identify new or changed segments.

While all of the above are good candidates for TM processing, they do have slightly different characteristics. Therefore, strictly at the level of fuzzy matching, some of these texts may be better candidates than others.

For instance, instinctively, it seems to me that a recycled text would be a case where a translator would need to pay careful attention to setting the fuzzy match value in order to maximise the benefit of the TM tool. Because recycled texts come from different sources and may be written by different authors, there is a lower

chance that exact matches will be found since everyone has their own idiolect which results in slight linguistic variation. Therefore, it may be necessary to choose a comparatively lower fuzzy match value in order to reap the benefits from the database. In contrast, I predict that in the case of a revision, the changes are likely to be less significant, and many passages will remain virtually untouched. Therefore, it is possible that a higher fuzzy match value will suffice.

I am not aware of any studies that have been conducted to investigate the impact that textual characteristics may have on fuzzy matching in particular, so I hope to go some way towards filling this gap by comparing the fuzzy match value needed to garner results from a TM database used to translate recycled texts as compared to the match value need to obtain similar results from a TM database used to translate a revised text.

2.3 The Language Factor

It is well known that different languages have different structural and stylistic preferences. For instance, when it comes to features such as cohesive devices, de Vries (2002: 46) notes that “where English tends to repeat words, in most other languages, referential pronouns are used.” De Vries (2002: 47) gives the following example, where the English sentences

*The installer can be downloaded from our site. You can use the
Installer to configure our software.*

would likely be translated into German as

Das Installationsprogramm kan von unserer Site heruntergeladen werden. Sie können es verwenden, um unserer Software zu konfigurieren..

De Vries notes that it would be possible to repeat *Installationsprogramm* instead of using the pronoun *es*, but it would compromise the readability of the text because the German word for “installer” is so much longer, and because this type of repetition is unnatural in German.

The real problem arises, however, because of the fact that a TM system works solely at the level of individual segments, rather than at the level of the text, so an antecedent contained in a different segment cannot be used to disambiguate the pronoun.

De Vries continues his example by asking what would happen if, in a subsequent document, the passage were updated to read:

The installer can be downloaded from our site. Here you will also find the product registration form. You can use the installer to configure our software.

The translator may need to provide a translation for the “new” (i.e., the second) sentence, but in principle, the first and third should be retrievable from the TM database. However, the fact that the pronoun *es* was used instead of repeating the noun *Installationsprogramm* has now resulted in a problem:

Das Installationsprogramm kan von unserer Site heruntergeladen werden. Hier finden Sie auch das Produktregistrierungsformular. Sie können es verwenden, um unserer Software zu konfigurieren.

As can be seen in the preceding translation, where the first and third sentences are taken from a TM database, it now appears to readers as if the registration form should be used to configure the software because *es* seems to refer back to *Produktregistrierungsformular* instead of to *Installationsprogramm*.

The point that can be taken from de Vries's example is that some languages may be more suited to being processed with the help of a TM system than others.

Mogensen (2000: 28) points out another difference between languages, this time between English and Danish. She notes that in technical texts in English, it is more common to find a series of short sentences, as compared to longer more complex structures, such as a main clause accompanied by subordinate clauses, which would be more typical of Danish technical texts. Again, the fact that a TM looks for repetition, which is more likely to occur in a shorter string rather than a longer one, would seem to indicate that some languages might be more conducive to translation with the help of a TM than others.

If we accept that different languages do have differing degrees of suitability with regard to TM processing, it raises the question as to whether different languages might therefore benefit from using different fuzzy match values. For instance, if we accept, as suggested by Vinay and Darbelnet (1958: 103, 109, 252) that English seems to use shorter sentences and repeat more words, whereas French uses longer constructions and prefers synonymy to repetition, it might therefore be the case that a lower fuzzy match threshold is needed when working with a French ST in

order to produce comparable results to those which can be obtained using a higher fuzzy match threshold when working with an English ST.

To the best of my knowledge, no one has conducted any specific investigations as to whether or not the languages or language direction have an impact on the fuzzy match setting, so as part of my research project, I will conduct a preliminary exploration of this question by testing my hypothesis that for the same type of text/TM database, translating from French to English will require the use of a lower fuzzy match threshold than will translating from English into French.

2.4 The Degree of Fuzziness Factor

It was noted in section 1.3.2 that many TM systems are capable of retrieving multiple fuzzy matches for a given segment. For example, if the match value is set to 75%, it is possible that the system might retrieve a 90% match, an 83% match and a 77% match.

The question that needs to be answered is whether or not the match with the highest percentage is always the best one? Macklovitch and Russell (2000: 138-39) and Bowker (2002b: 200) have pointed out that, because of the difference between semantic similarity and superficial similarity (see section 1.3.2), it is possible that the “best” match identified by the system may not actually be the most useful match for the translator. But how often does this really happen? To the best of our knowledge, no one has undertaken an empirical study of the multiple

matches suggested by a TM system with a view to evaluating how often the “best” match proposed by the system is actually the most helpful match for the translator. Somers (2003b: 43) hints at this issue when he notes that “adapting partial matches to the new input is not the same as translating from scratch, but *the amount of work may or may not be directly proportional to the percentage of fuzzy match on the source text*” [my emphasis]. In other words, Somers queries whether a higher percentage match would mean less work for a translator as compared to a lower percentage match. However, he does not actually provide a definitive answer to this question, offering only anecdotal evidence that “some informal research by this author ... has suggested that at the top end, the fuzzy match scores probably do give, for practical purposes, as good a measure of similarity with matches in the database as could be desired.” (Somers 2003b: 43)

Likewise, Höge (2002: 171) makes the assumption that “If the minimum match value in the system is set to 30%, more matches with presumably a lower quality will be retrieved than if it is set to 60%.” In other words, she assumes that a match with a lower percentage will be of a lower quality, and therefore implicitly less useful to the translator.

Instinctively, this reasoning seems logical, but no one appears to have undertaken any serious empirical testing, which accounts for all the hedges and vague statements (e.g., “suggested,” “probably,” “presumably”) that appear in the literature. Therefore, the belief that a higher percentage match is usually a better match remains, at the present time, an axiom rather than a truth. The Gage

Canadian Dictionary defines *axiom* as “a statement seen to be true without proof.”

Therefore, as part of this research project, I will attempt to verify the axiom that a higher percentage match is typically more helpful to a translator. Based on the results of this investigation, I will be able to determine whether what is currently an axiom can actually be considered as fact.

2.5 Concluding remarks

The four research questions pertaining to fuzzy matching that have been outlined in this chapter will now be investigated in a series of empirical studies described in the following four chapters.

3 The System Factor

As noted in section 2.1, one of the research questions that will be investigated in the context of this thesis is the question of the extent to which a given TM system affects the results of the fuzzy matching process. This chapter will set forth the methodology that was used to compare fuzzy matches generated by two popular TM systems. The general objective of this investigation into the “System Factor” is to determine whether both systems retrieve the same set of fuzzy matches from an identical TM database, and in cases where the same data is retrieved, whether one system produces a different fuzzy match score than the other.

3.1 TM Tool Selection

It was pointed out in section 1.1 that there are many competing TM systems on the market. Ideally, it would be useful to compare all the available systems in order to draw up a comparison table of fuzzy match settings that translators could consult when they set out to use a given system, particularly if they are not very familiar with that system. However, in the context of an MA thesis, where time and resources are limited, it was necessary for me to restrict this investigation to two systems: Trados Translator’s Workbench and Fusion Translate. However, it is important to note that the methodology used to compare the results from these two systems is generic and could in fact be applied to any system.

As noted in section 0.4.1, the primary motivation for selecting Trados and Fusion Translate is that both are available in the “Writing Centre” computer laboratory operated by the Faculty of Arts at the University of Ottawa.

3.2 TM Database Design and Construction

After selecting the two tools to be used in this investigation, I next set out to design and compile TM databases that could be processed by each tool. A detailed discussion of texts that are suitable for TM processing will be given in Chapter 4: The Text Category Factor. However, I will provide a brief description of the texts used to create the TM databases here, along with a description of the database compilation process.

Researchers such as Webb (1998) and Bowker (2002a) have observed that text types that are formulaic in nature (e.g., texts that follow a specific format, that use set expressions, and that have field specific terminology) make good candidates for use with TM systems. As described in section 1.2.2.5.3, the goal of a TM system is to recycle portions of previously translated texts. It therefore makes sense that texts that are formulaic in nature will be good candidates for this type of recycling. In addition to identifying formulaic texts, I also needed to find material that had been previously translated and that was available in English and French so that I could align and store this material in a TM database.

3.2.1 TM Database 1: Appeal Decisions

Since material on the Government of Canada's websites is readily available in both English and French, I began my search there. Moreover, since I am currently in the process of becoming a Canadian citizen, I have become somewhat familiar with the Citizenship and Immigration Canada (CIC) website¹², and an exploration of this site eventually led me to discover a searchable database called *RefLex*¹³ which contains summaries of immigration appeal decisions, along with links to the full-text decisions in both of Canada's official languages.

An examination of the appeal decisions revealed that they were quite formulaic in nature. Each decision is on average three to five pages in length and contains sections such as "Background," "Analysis" and "Notice of Decision." While the information in the background section varies according to the nature of each case, there are a limited number of decisions that can be reached (i.e., immigration appeals can be either denied or allowed) and the legal arguments used as part of the analysis are often quite similar. In addition, the terminology that is part of this specialised field also appeared on a regular basis in these documents.

Having decided that the appeal decisions were a good candidate for processing with the help of a TM system, I then began to build databases for both Trados and Fusion. Because these two tools store data in a proprietary format, it was not possible to create a single database for use with both tools. However, it was imperative that

¹² The CIC website is located at: <http://www.cic.gc.ca/>

¹³ http://www.irb-cisr.gc.ca/en/decisions/reflex/index_e.htm?action=search

the contents of both databases be exactly the same. I began the database construction by going to the *RefLex* database and searching for appeal decisions. The search interface for *RefLex* allows users to restrict searches by division, language, date and keyword. In order to maximise recyclability of the texts, I began searching in English and limited my search to those decisions where the appeal had been allowed. In addition, I focused on the time period 1997 to 2005 since the most recent decisions are not yet available in both languages.

After conducting a search in *RefLex*, the user is presented with a list of documents (shown as hyperlinks) that meet the search criteria. Clicking on the document name brings the user to a summary of the decision, and another click is required to pull up the complete decision, which opens as a Word document. This document can then be saved on the user's hard disk. By returning to the summary page, the user can select to view the corresponding French-language summary, and from there, open the Word document containing the complete decision in French. This, too, can be saved on the user's hard disk. This process must be repeated for each text to be included in the database. Given that it is both labour-intensive and time-consuming, I restricted the size of my appeal decisions database to 15 decisions (in both English and French), for a total of 149 translation units in Trados and 151 translation units in Fusion¹⁴ (see section 1.2.1.3).

¹⁴ Note that because each TM system has its own proprietary alignment tool, the total number of TUs may differ slightly. For example, one system may consider a semicolon to be a marker that indicates a division between two units, whereas the other TM system may consider clauses joined by a semicolon to belong to the same unit.

Once I had the Word documents, I was able to use these to generate the required proprietary format TM databases in both Trados and Fusion Translate. However, since automatic alignment tools are not perfect (see section 1.2.1.2), it was necessary to spend time (approximately 20 hours) manually verifying and correcting the alignment in both TM systems. This was necessary because if the alignment is not correct, the supposed “matches” retrieved from the TM database will not actually be accurate.

3.2.2 TM Database 2: Translation Job Advertisements

Having constructed one TM database, I felt it was important to compile a second database using different texts. I made this decision because I felt it would help me to reach conclusions about the “System Factor” that were more generalizable. In other words, if I used only a single TM database, I could not be sure whether this database just happened to be more suitable for one system than the other. By using a second database, I would be able to see if each system behaved in a similar way when processing both databases.

For my second database, I chose another type of formulaic text: job advertisements. Because job ads typically contain similar sections (e.g. duties, qualifications, starting date, salary) I determined that they would also be a good candidate for recycling information with the help of a TM. In order to maximise the recyclability, I restricted the domain of the job ads to language professionals (e.g., translators, revisers, editors, terminologists) so that the terminology of the field would also be repeated.

I was able to locate 16 job ads (eight in English, and their eight translations in French) for language professionals distributed by the Association of Translators and Interpreters of Ontario (ATIO). Some of these ads were already in Word format, while others were in HTML format, but their contents could easily be copied and pasted into a Word document. It was necessary to separate the bilingual contents of the ads into two separate files – the English version of the ad in one file, and the French version in another – because the automatic alignment tools in both Trados and Fusion Translate require the texts to be in separate files.

To create the TM databases, I followed the same procedure as described above for the appeal decisions, including automatic alignment and manual verification. The end result was two databases (one for Trados and one for Fusion Translate) containing identical contents in the proprietary format required by each tool. When stored in the databases, these job ads databases contained 149 TUs in Trados and 155 TUs in Fusion¹⁵. This is very comparable to the number of TUs contained in the appeal decisions database.

3.3 Source Text Selection

As described in section 1.2.2, a TM system allows a translator to compare the contents of a new source text (or texts) against the contents of the TM database in order to identify any sections of the source text that have previously been translated.

¹⁵ Once again, because each TM system has its own proprietary alignment tool, the total number of TUs may differ slightly even though the exact same texts are contained in both databases.

I therefore needed to find some new source texts to compare against the TM database contents.

For the appeal decisions, I simply returned to the *RefLex* database and extracted 20 additional decisions that had not been included in the TM database. I chose only the French-language versions of these decisions since, as an Anglophone, I translate normally from French into English. For the job advertisements, I selected nine additional job ads for language professionals that had been advertised in French only by the ATIO.

3.4 Testing Procedure

The first step in the testing procedure was to set the minimum fuzzy matching value threshold to the lowest setting possible in each system, which corresponds to 30% in Trados, and 50% in Fusion Translate. This was done in order to generate the maximum possible number of fuzzy matches for data comparison.

Next, each of the new source texts to be translated was put through both TM systems, and the resulting fuzzy matches generated from the TM databases were recorded. Because the focus of this study is on fuzzy matches, other types of matches (e.g., exact, sub-segment and term matches) were not considered. The detailed results produced by each system and TM database can be found in Appendices 3.1 and 3.2.

3.5 Data Analysis

The results will be analyzed and described below. For the sake of clarity, I will first discuss the results of the appeal decisions TM database, comparing the findings for Trados against those for Fusion Translate. Next, a similar discussion will focus on the results of the job advertisements TM database. In both cases, the results will be discussed from both a quantitative and a qualitative viewpoint.

3.5.1 Appeal Decisions TM Database

Table 3.1 presents a comparative summary of the scores for the fuzzy matches retrieved by each system from the appeal decisions TM database. Numbers shown in square brackets indicate that a second fuzzy match was retrieved for a given segment. The right-hand column represents the difference in performance of the two TM systems, expressed in relation to Trados' performance (i.e., +3% means that the best fuzzy match found by Trados scored 3% higher than that found by Fusion Translate).

New segment to be translated	Trados scores	Fusion scores	Difference (expressed in relation to Trados' performance)
Segment 1	35%	No match found	
Segment 2	61%	58%	+3%
Segment 3	35%	No match found	
Segment 4	45%	No match found	
Segment 5	39%	No match found	
Segment 6	65%	61%	+4%
Segment 7	39%	No match found	
Segment 8	59%	No match found	
Segment 9	67%	53%	+14%
Segment 10	37%	53%	-16%
Segment 11	82% [64%]	72%	+10%
Segment 12	67% [57%]	55% [55%]	+12%
Segment 13	41%	63%	-22%
Segment 14	65%	No match found	
Segment 15	87% [80%]	83%	+4%
Segment 16	57%	100%	-43%
Segment 17	74% [64%]	69%	+5%
Segment 18	59% [58%]	No match found	

Table 3.1 – Comparative Analysis of TM Tools – Immigration Appeals TM Database.

3.5.1.1 Quantitative Analysis

This section presents the results of the quantitative analysis for the appeal decisions TM database.

3.5.1.1.1 Lack of Fuzzy Matches in Fusion

As mentioned in section 3.4, the fuzzy match threshold was set to the lowest possible value in each system. However, this value is not the same for both systems. In Trados, the lowest possible fuzzy match threshold is 30%, while that of Fusion is considerably higher at 50%.

Looking at Table 3.1 above, it seems that Trados outperforms Fusion with regard to the overall number of fuzzy matches retrieved since in 8 out of 18 cases, Fusion retrieved no match whatsoever. However, if we examine those 8 instances more closely, we can see that in 5 out of 8 cases (i.e., segments 1, 3, 4, 5 and 7), Trados retrieved a match with a score that was less than 50%, so perhaps it is not surprising that Fusion did not retrieve matches for those cases. If it were possible to set the match value in Fusion down to 30%, then Fusion may also have retrieved matches for those segments.

What is quite surprising, however, is the fact that Fusion did not generate matches for segment 8 (where Trados had a score of 59%), segment 14 (where Trados scored 65%) and segment 18 (where Trados retrieved two fuzzy matches with scores of 59% and 58% respectively). If we consider the instances where both Fusion and Trados retrieved a fuzzy match, but where Trados received a higher score, we can see that these scores range from being between 3% and 14% higher. Therefore, at the very least, we might have expected Fusion to retrieve a match for segment 14, since the Trados match was, at 65%, a value that was 15% higher than Fusion's minimum match value of 50%.

3.5.1.1.2 Disparity in Scores

In the 10 out of 18 cases where *both* Trados and Fusion retrieved a fuzzy match¹⁶ (see Table 3.1), Trados outscored Fusion in 7 out of those 10 cases, whereas Fusion outscored Trados in only 3 out of those 10 instances. However, the margin by which Fusion outscored Trados was significant (ranging from 16%-43%) whereas when Trados outperformed Fusion, it was typically by a much smaller range (3%-14%). In 4 of the 7 cases where Trados outperformed Fusion, the difference was 5% or less.

However, looking at the data in a slightly different way, Table 3.2 presents those instances in which the *same* segment was retrieved by both TM systems. It is interesting to note that the disparity between the fuzzy matching results is not as significant in these cases. Fusion generated a higher percentage of fuzzy matches in 4 out of 10 matches, whereas Trados did slightly better, generating 6 out of 10. However, the average difference in favour of Trados was only 6.5%, whereas when Fusion outperformed Trados, it was by an average of 12.8%.

¹⁶ It is important to note here that although both systems did in fact retrieve at least one fuzzy match from the TM database, it was not necessarily the case that both systems retrieved the *same* segment. A comparison of cases where the same segment was retrieved will be presented further on.

New segment to be translated	Trados	Fusion	Difference (expressed in relation to Trados' performance)
Segment 2	61%	58%	+3%
Segment 6	65%	61%	+4%
Segment 9	67%	53%	+14%
Segment 10	37%	53%	-16%
Segment 11	64%	72%	-8%
Segment 12	67%	55%	+12%
Segment 12a) ¹⁷	57%	55%	+2%
Segment 13	41%	63%	-22%
Segment 15	87%	83%	+4%
Segment 16	64%	69%	-5%

Table 3.2 – Instances where the same segment was retrieved by both systems – appeal decisions

3.5.1.1.3 Disparity with Regard to Multiple Matches

Another result that merits special discussion is the situation where multiple fuzzy matches were retrieved for a given segment. Take the case of Segment 15, for example. The Trados system generated two relatively high fuzzy matches for this segment (87% and 80% respectively), whereas Fusion only presented one fuzzy match at 83%, which corresponded to Trados's 87% match. A somewhat similar situation arose in the case of Segment 17, where Trados again generated two matches with moderately high scores (74% and 64%), whereas Fusion generated only one match. However, the Fusion match, which had a score of 69%, corresponded to the same match that Trados had retrieved with a 64% score. In other words, the segment that Trados retrieved with a 74% match was not retrieved at all by Fusion. Segment 11 produced a similar kind of result, with Trados generating 2 matches ((82% and 64% respectively), while Fusion had only a single match of 72%, which corresponded to the 64% match generated by Trados.

¹⁷ For segment 12, both systems retrieved two potential matches. 12a) represents the 2nd match retrieved by both systems.

For the 18 segments that generated fuzzy matches, there were 5 cases where Trados offered the translator multiple fuzzy matches, whereas there was only a single instance where Fusion generated multiple fuzzy matches (Segment 12), and interestingly, these both had exactly the same fuzzy matching value (55%).

3.5.1.2 Qualitative Analysis

In an effort to better understand why the two TM systems produced the results they did, a qualitative examination of the data was performed. Although it is not conclusive, it may nonetheless shed some light on the situation and will hopefully inspire further qualitative studies into fuzzy matching results.

3.5.1.2.1 Type of Match and Noise Level

The first segments that I examined were the 8 segments for which Trados found a match but Fusion did not. In examining these segments, it soon became clear that the match was being generated largely on the basis of a consecutive sub-segment string, as shown in the examples in Table 3.3. Apart from that sub-segment string, the two segments had relatively little in common.

Segment 1	New ST segment: La comparution en personne n'est pas accordée sur simple demande et, aux termes du paragraphe 175(2) <u>de la Loi sur l'immigration et la protection des réfugiés</u>, il doit y avoir des motifs qui étayent une telle demande. Stored segment: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 <u>de la Loi sur l'immigration et la protection des réfugiés</u>.
Segment 3	New ST segment: <u>La demande parrainée de résidence permanente de la partenaire conjugale de l'appelant</u> a été refusée. Stored segment: L'appelant en a appelé du refus de <u>la demande parrainée de résidence permanente de sa femme</u>.
Segment 8	New ST segment: Lors de l'appel, <u>l'appelant n'a pas contesté la validité</u> de cette mesure. Stored segment: <u>L'appelant n'a pas contesté la validité</u> juridique du refus.

Table 3.3 – Examples of fuzzy matches generated by Trados that were largely the result of a single sub-segment string match.

As noted in section 1.2.2.3, a number of TM systems are now making use of a different type of matching known as sub-segment matching. Whereas Trados places a significant emphasis on fuzzy matching, Fusion Translate is a tool that makes heavy use of sub-segment matching. Therefore, I felt it was highly likely that in Fusion, these particular matches would in fact be presented as “expression matches”¹⁸ rather than as fuzzy matches. A quick investigation proved this to be correct – these sub-segments were all retrieved by Fusion in the form of “expression matches” rather than as fuzzy matches.

¹⁸ “Expression match” is the terminology used by Fusion Translation to refer to a sub-segment match.

3.5.1.2.2 Different scores for same segments

In those cases where both systems retrieved the same segment, but assigned a different score, I could discern no underlying pattern to explain this phenomenon. Without having access to the actual algorithms used by the two systems to generate fuzzy matches, it is not possible to fully understand the criteria used to assign a score to a fuzzy match.

3.5.2 Job Advertisement TM Database

As was the case for the appeal decisions database above, Table 3.4 presents a comparative summary of the scores for the fuzzy matches retrieved by each system from the job ads TM database. Numbers shown in square brackets indicate that a second fuzzy match was retrieved for a given segment. The right-hand column represents the difference in performance, expressed in relation to Trados' performance (i.e., +2% means that the best fuzzy match found by Trados scored 2% higher than that found by Fusion Translate).

New segment to be translated	Trados scores	Fusion scores	Difference (expressed in relation to Trados' performance)
Segment 1	36% [33%]	No match found	
Segment 2	69%	No match found	
Segment 3	75%	73%	+2%
Segment 4	46%	100%	-54%
Segment 5	36% [35%]	No match found	
Segment 6	34%	No match found	
Segment 7	60% [56%] [38%]	55%	+5%
Segment 8	52% [52%]	No match found	
Segment 9	43%	No match found	
Segment 10	42%	No match found	

Segment 11	37%	No match found	
Segment 12	44%	No match found	
Segment 13	35%	No match found	
Segment 14	37%	No match found	
Segment 15	56%	67%	-11%
Segment 16	100%	68%	+32%

Table 3.4 – Comparative Analysis of TM Tools – Job Advertisements TM Database

3.5.2.1 Quantitative Analysis

This section presents the results of the quantitative analysis for the appeal decisions TM database.

3.5.2.1.1 Lack of Fuzzy Matches in Fusion

The results in Table 3.4 show that once again, Trados appears to significantly outperform Fusion given that there were 11 out of 16 instances where Fusion did not find any match at all. However, keeping in mind that the Trados minimum match value was set to 30% while the Fusion match value was set to 50%, we can see that in 9 of the 11 cases where Fusion did not find a match, the score generated by Trados was less than 50%, and in the tenth case, the Trados score was only 52%. Given the difference in the minimum match value settings, it seems reasonable that Fusion did not find matches in those 10 cases; however, it is surprising that Fusion was not able to generate a fuzzy match for segment 2, which Trados scored at 69%.

3.5.2.1.2 Disparity in Scores

In the 5 out of 16 cases where *both* Trados and Fusion retrieved a fuzzy match (see Table 3.5) Trados outscored Fusion in 3 out of those 5 cases, whereas Fusion outscored Trados in only 2 out of those 5 instances. With regard to the margin, on 2 of the 3 occasions where Trados had the better score, it was by a margin of 5% or less, whereas the third time it was by a more significant score of 32%. Meanwhile, on the 2 occasions where Fusion outperformed Trados, it was once again by a more significant margin: by 11% in one instance and by 54% in the other.

However, looking at the data in a slightly different way, Table 3.5 presents those instances in which the *same* segment was retrieved by both TM systems (shown with a check mark symbol (✓)), as well as those occasions where both systems retrieved a fuzzy match, but it was *not* the same segment that was retrieved (shown with an X).

New segment to be translated	Trados scores	Fusion scores	Difference (expressed in relation to Trados' performance)
Segment 3 ✓	75%	73%	+2%
Segment 4 X	46%	100%	-54%
Segment 7 ✓	60%	55%	+5%
Segment 15 X	56%	67%	-11%
Segment 16 X	100%	68%	+32%

Table 3.5 - Instances in which the same segment was retrieved by both TM systems

In those cases where the very same segment was retrieved by both systems (i.e., segments 3 and 7) the difference in performance is 5% or less. In contrast, where the two systems retrieve completely different segments as matches, the difference is over 10%.

3.5.2.1.3 Disparity with Regard to Multiple Matches

As was the case with the appeal decisions database, the job ads database also presented differences with regard to the number of multiple matches generated. Once again, Trados outperformed Fusion in this regard, with multiple matches being generated on four occasions, as compared to no instances of multiple matches being generated by Fusion. However, it should be noted that on 2 of the 4 occasions where Trados generated a multiple match, the score of the second match was below 50%, which is below the minimum match threshold allowed by Fusion.

3.5.2.2 Qualitative Analysis

In order to better understand the results of the quantitative study, this section will analyse the data from a qualitative viewpoint.

3.5.2.2.1 Type of Match

The first segments that I examined were the 11 segments for which Trados found a match but Fusion did not. This investigation revealed results that were similar to the observations made for the appeal decisions database: the matches in Trados were being generated largely on the basis of a consecutive sub-segment string, as shown in the examples in Table 3.6.

Segment 1	New ST segment: Vous vous intéressez à l'économie et à la finance et comptez <u>au moins cinq années d'expérience comme traducteur dans</u> ces domaines ou d'autres domaines pertinents? Stored segment: Vous avez <u>au moins deux</u>¹⁹ <u>années d'expérience comme traducteur dans</u> un ou plusieurs des domaines suivants :
Segment 13	New ST segment: Toutes les candidatures seront traitées confidentiellement et dans le respect du principe d'<u>équité en matière d'emploi</u>. Stored segment: Nous souscrivons au principe de l'<u>équité en matière d'emploi</u>.
Segment 14	New ST segment: Si ce défi vous intéresse, vous pouvez acheminer votre curriculum vitae selon votre convenance à l'adresse courriel : Stored segment: Pour soumettre votre candidature à ce poste, veuillez faire parvenir votre curriculum vitae d'ici le 8 août, 2003 à l'adresse suivante :

Table 3.6 – Examples of fuzzy matches generated by Trados that were largely the result of a single sub-segment string match.

Once again, a quick investigation into the “expression match” feature of Fusion revealed that although these matches were not being retrieved as fuzzy matches, the information was in fact being presented to the translator, but as a sub-segment match rather than as a fuzzy match.

3.5.2.2.2 Different scores for same segments

Unfortunately, given that there were only two instances where both Trados and Fusion retrieved the same segment, there was not enough data to generate any interesting qualitative analyses.

3.6 Conclusions

In looking at the results of the analysis of both the appeal decisions database and the job advertisements database, we can see that the two investigations generated

¹⁹ Note that TM systems typically ignore so-called “variable elements” such as numbers when making a match between segments (Macklovitch and Russell 2000).

very similar findings, which means that our conclusions are generalizable (to a limited degree).

On the basis of a purely quantitative analysis of the fuzzy matching performance of the two systems, one has the impression that Trados has “more” to offer the translator. In both cases, Trados generated matches for a greater number of segments (18 for Trados vs. 8 for Fusion for the appeal decisions database, and 16 for Trados vs. 5 for Fusion for the job ads database). Moreover, Trados frequently generated multiple possible matches for a given segment (5 multiple matches in the appeal decisions database and 4 in the job ads database), whereas Fusion did so on only a single occasion for the appeal decisions database, and not at all for the job ads database. In addition, in instances where both systems generated a match, those proposed by Trados more often than not had a higher score than those suggested by Fusion. However, when Fusion did generate matches that scored higher than those of Trados, the margin was usually more significant.

Nevertheless, the results of the qualitative analysis attenuate these quantitative findings somewhat, showing that in many cases Fusion does not actually “miss” matches, but rather simply categorises them differently (i.e., as “expression matches” rather than as fuzzy matches). What’s more, these expression matches may actually be easier to consult than the low percentage fuzzy matches retrieved by Trados because the Trados fuzzy matches that had low scores (e.g., between 30% and 50%) contain a lot of extraneous noise. Therefore, while at first glance Fusion seemed to show an inferior performance the opposite may actually be true.

We could say that although Fusion produced fewer fuzzy matches, the fuzzy matches that it did produce were of a higher quality and therefore more helpful. Meanwhile the information that at first glance seemed to be overlooked by Fusion has in fact been recognised and is merely shown as a different category of match (i.e., a sub-segment match rather than a fuzzy match).

This raises the point that any meaningful comparative evaluation of different TM systems must be carried out at a more global level and should take into account all the features of these systems, and not just a single feature, such as fuzzy matching. Nevertheless, since fuzzy matching is one of the important features of a TM tool, it is still useful for translators to have an understanding of TM system performance in this regard. Based on my investigation, I can summarise the following “System Factor” points as being of key importance to a translator comparing Trados with Fusion Translate in terms of fuzzy matching:

- Trados will retrieve a greater number of fuzzy matches than Fusion.
- Many of the fuzzy matches that are retrieved by Trados but not by Fusion are low quality, noisy matches.
- Many of the low quality, noisy matches that are retrieved as fuzzy matches by Trados will be retrieved as sub-segment matches by Fusion, so translators working with Fusion need to be aware that they should check the sub-segment matches for information not retrieved in the form of fuzzy matches.
- Trados is much more likely than Fusion to retrieve multiple fuzzy matches for a given segment. Since information that is useful to a translator may not

always be contained in the segment with the highest match percentage, Trados would seem to offer a greater possibility to translators to find useful information.²⁰

- When Trados and Fusion both retrieve the same segments for a fuzzy match, the majority of the time, Trados gives a score that is between 2% and 5% higher than that assigned by Fusion. Therefore, translators working with Fusion would need to set their minimum match value approximately 5% lower than they would when working with Trados in order to achieve comparable results. For example, a minimum match value setting of 75% in Trados would generate results comparable to a setting of approximately 70% in Fusion. A translator who is used to working with Trados and who is asked to switch to Fusion might inadvertently miss a number of useful matches if he or she chose to use a 75% setting (if that is the setting that he or she normally uses with Trados).

The results of this small investigation demonstrate that the “System Factor” is definitely in play when it comes to setting the fuzzy match threshold in a TM system. Different systems processing the same set of data will produce different results at the same setting. Clearly the findings presented here are rather limited since this investigation compared only two TM systems. Ideally, a comparative analysis of a wider range of TM systems using a similar type of methodology would be helpful in generating data and empirical evidence that would give translators a better idea of what they might expect when using a TM tool not familiar to them and would help

²⁰ This point will be discussed in greater detail in Chapter 6: The Degree of Fuzziness Factor.

them to set the fuzzy matching value percentage at the level where it would produce the most useful results. Nevertheless, the study presented here has at least demonstrated that further investigations into the “System Factor” are merited.

In the following chapter, I will explore another issue, which I refer to as the “Text Category Factor,” to see if this has any impact on fuzzy matching in TMs.

4 The Text Category Factor

Another issue that will be investigated as part of this thesis is the question of whether the nature of the texts that are processed with the help of a TM system has an impact on fuzzy matching. In this chapter, I will explore the “Text Category Factor” by determining whether texts that possess different characteristics require users to set the fuzzy match values at different levels. The following sections present the methodology used in this investigation, as well as an analysis of the results, including a set of preliminary guidelines that users can keep in mind when setting the fuzzy match value for different categories of text.

4.1 Text Selection

In section 2.2, four categories of text were identified as being reasonably well-suited for processing with the help of a TM system. These categories were:

- **Texts containing internal repetitions:** Texts containing a great deal of repetitive content.
- **Recycled texts:** Texts that come from the same client/subject field/genre as previously written (and translated) texts.
- **Revisions:** Texts that are amended versions of previously completed (and translated) texts.
- **Updates:** Texts where changes are made to the source text while the translation is still in progress.

Based on the information gleaned from the literature about what categories of texts can most usefully be translated with the help of a TM system, I set out to identify some texts that could be used as part of this research project. Because it was my intention to try and build a TM database containing multiple texts, I decided not to consider the category “texts containing internal repetitions” since this focused on the potential for recycling material within a single text.

Likewise, since I was not working for a client who would be providing updates, I decided not to pursue this category either. Instead, I decided to try and identify texts that could be used to build two different databases — one for the category of “Recycled Texts” and the other for the category “Revisions.” I felt that of the four categories described by Webb (1998) and Bowker (2002a), these were the two that translators and TM users would most likely encounter. By creating TM databases for these two different categories of texts, each of which have different characteristics, I hoped to determine whether the “Text Category Factor” comes into play during fuzzy matching.

In addition to identifying texts from the appropriate categories (i.e., “recycled texts” or “revisions”), I also had to make sure that these texts were available in English and French, which are the working languages of this project.

4.1.1 Recycled Texts

According to the description provided by Bowker (2002a: 113-114), the main characteristics for the category of “recycled texts” are that the texts should come from the same subject field, genre and/or client. In this way, the texts are likely to contain similar terminology and structures that can be recycled. Looking back to the two TM databases that I had already constructed for the test carried out in Chapter 3, I determined that both the appeal decisions database and the job advertisements database fell into this category of “recycled texts.” The appeal decisions database contains appeal cases heard by the Immigration and Refugee Board that were either allowed or denied, while the job advertisement database contains a variety of advertisements seeking language professionals. As described in section 3.2, both databases contain texts that have a formulaic structure and a restricted terminology, so both seemed to be good examples of material fitting into the “recycled texts” category. However, since the results of the experiment in Chapter 3 showed that the appeal decision database generated a higher number fuzzy matches as compared to the job ads database, I decided to retain the former as one of the databases to use in the “Text Category Factor” experiment. I also retained the new source text described in section 3.3 for use in this experiment.

4.1.2 Revisions

Since I had not constructed any databases containing revised texts for the work carried out in Chapter 3, it was necessary for me to create a new TM database for this category.

My supervisor informed me that the Association of Professors of the University of Ottawa (APUO) was in the process of re-negotiating their collective agreement, so I decided to investigate this as a potential source of material for the database of revised texts. The complete collective agreement for the period 1 May 2001 to 30 April 30 2004 was available online in both French and English at the APUO website²¹. I was able to download this document and use it to generate a TM database following the same procedure described in section 3.2. This database contained 3,517 translation units (TUs).

For the new source text to be translated, I was able to obtain a copy of the document entitled “Entente Proposée, Convention Collective, 1 mai 2004 au 30 avril 2008 / Tentative Settlement, Collective Agreement, 1 May 2004 to 30 April 2008.” This document contained the proposed changes that would be made to the APUO collective agreement. The following statement also appeared on the cover of this document: “Les textes sont présentés seulement dans la langue dans laquelle ils ont été finalement négociés. / Texts are presented only in the language in which they were finally negotiated.” This meant that some of the sections of the new text were in English only and needed to be translated into French, while others were in French only and needed to be translated into English. For this part of the experiment, I selected only the French-language sections that needed to be translated into English since this is the language direction in which I normally work.

²¹ APUO website: <http://www.apuo.uottawa.ca/Info/Convention/0104-CONT.htm>

4.2 TM System Selection

To proceed with the experiment to test the “Text Category Factor,” only one TM system was needed. I opted to use the Trados TM tool since this had generated a greater number of fuzzy matches during the “System Factor” experiment described in Chapter 3.

4.3 Testing Procedure

Once again, the lowest possible fuzzy match setting in Trados (i.e., 30%) was used in order to maximise the number of fuzzy matches retrieved. Each of the new source texts to be translated was processed using the corresponding TM database. In other words, the appeal decisions source text was processed using the appeal decisions database, and a number of fuzzy matches were generated, while the text containing the APUO collective agreement revisions was processed using the APUO collective agreement database and a number of fuzzy matches resulted. The fuzzy matches were recorded and were used as the basis for a comparative evaluation of the “Text Category Factor.” The detailed results for each of the two TM databases can be viewed in Appendix 4.1.

4.4 Test Results and Data Analysis

Table 4.1 below presents the results of fuzzy matches generated for the “Recycled Texts” and Table 4.2 for the “Revisions.” For the purpose of this experiment, I focused on the “best” (i.e., highest scoring) fuzzy match produced for each segment.

Recycled Texts (appeal decisions)	Fuzzy match score (%)
Segment 1	35%
Segment 2	61%
Segment 3	35%
Segment 4	45%
Segment 5	39%
Segment 6	65%
Segment 7	39%
Segment 8	59%
Segment 9	67%
Segment 10	37%
Segment 11	82%
Segment 12	67%
Segment 13	41%
Segment 14	65%
Segment 15	87%
Segment 16	57%
Segment 17	74%
Segment 18	59%
TOTAL No. of matches	18
Average score of fuzzy matches	56%
No. of fuzzy matches scoring 80% or higher	2 (= 11%)
No. of fuzzy matches scoring 50% or lower	7 (= 39%)

Table 4.1 – Results for “Recycled Texts” (appeal decisions)

Revisions (APUO collective agreement)	Fuzzy match score (%)
Segment 1	83%
Segment 2	85%
Segment 3	36%
Segment 4	33%
Segment 5	91%
Segment 6	64%
TOTAL No. of fuzzy matches	6
Average score of fuzzy matches	65%
No. of fuzzy matches scoring 80% or higher	3 (= 50%)
No. of fuzzy matches scoring 50% or lower	2 (= 33%)

Table 4.2 – Results for “Revisions” (APUO collective agreement)

4.4.1 Number of Matches Retrieved

With regard to the sheer quantity of matches retrieved, the “Recycled Texts” produced a greater number of fuzzy matches. In this case, a total of 18 fuzzy matches were generated for the “Recycled Texts,” whereas a total of 6 fuzzy matches were generated in the “Revisions” category. However, we must keep in mind that the source text used in the “Revisions” category was considerably shorter than the one used in the “Recycled Texts” category because there were a limited number of changes that were made to the collective agreement during the bargaining process.

4.4.2 Average Score of Fuzzy Matches Retrieved

By calculating the mean score of the fuzzy matches retrieved, we get an average of 56% for the “Recycled Texts,” as compared to an average of 65% for the “Revisions.” This means that, on average, the fuzzy matches generated by ‘Revisions’ are close to 10% higher than those produced by “Recycled Texts.”

4.4.3 Number of High Quality and Low Quality Matches Retrieved

Although the “Recycled Texts” generated a greater number of fuzzy matches, the “Revisions” produced a much more significant ratio of high quality fuzzy matches (where high quality is considered to be a match scoring 80% or greater). In the case of the “Revisions,” 3 out of 6 (or 50%) of the fuzzy matches were high quality matches, whereas for the “Recycled Texts,” only 2 out 18 (or 11%) of the fuzzy matches qualified as high quality.

In a similar vein, the “Revisions” generated a somewhat smaller number of poor quality fuzzy matches (where poor quality is understood to be a match scoring 50% or less). The number of poor quality matches produced for the “Revisions” was 2 out of 6 (or 33%), while the “Recycled Texts” generated a total of 7 out of 18 (or 39%) fuzzy matches that were deemed to be of poor quality.

4.5 Conclusions

Upon examining the results of the analysis of both the “Recycled Texts” and the “Revisions,” we can see that both categories of text are good candidates for use with a TM system, but for different reasons.

The “Recycled Texts” are good candidates because they generate a large number of fuzzy matches. This means that translators will have lots of material that they can consult, and they can possibly find helpful solutions to assist them with their translation problems. The downside of this is that relatively few of these matches appear to be of extremely high quality. In contrast, the “Revisions” are good candidates because, although they generate fewer fuzzy matches, those that they do generate are typically of quite a good quality and will therefore likely be of assistance to the translator.

When it comes to setting the fuzzy match value in a system, translators who are processing “Recycled Texts” might therefore wish to set the fuzzy match threshold

to a lower point than they would if they were translating “Revisions.” For example, based on the data presented in section 4.4 above, we can see that in the case of the “Recycled Texts,” a match value set at 60% would retrieve 8 out of 18 (or 44%) of the possible fuzzy matches. In contrast, the same 60% match value would retrieve a total of 4 out of 6 (or 66%) of the possible fuzzy matches for the “Revisions.” To retrieve 66% of the possible fuzzy matches from the “Recycled Texts,” the translator would need to set the fuzzy match value to 45%. Similarly, calculating the mean fuzzy match score for both “Recycled Texts” and “Revisions” shows that the scores produced in the “Revisions” category are, on average, nearly 10% higher than those produced by the “Revised Texts.”

Based on this investigation of the “Text Category Factor,” I can therefore summarise the following points as being of importance to a translator with regard to fuzzy matching and the processing of different categories of text:

- “Recycled Texts” yield a greater number of fuzzy matches; however, many of these are of a relatively low quality.
- “Revisions” generate a smaller number of fuzzy matches, but these are of a higher quality as evidenced by the higher percentages of the scores.
- Translators working with “Recycled Texts” need to use a lower fuzzy match threshold than translators working with “Revised Texts” in order to retrieve a comparable fraction of fuzzy matches.

The results of this small investigation would seem to indicate that the “Text Category Factor” is something that translators should consider when it comes to setting the fuzzy match threshold in a TM system. Of course, the findings presented here are quite limited since they are based on the results of a study that involved only two texts with a relatively small number of fuzzy matches. To produce more definitive results, it would be necessary to use a larger number of texts from both of the categories in question. However, this study has at least demonstrated that the methodology used seems viable, and that further investigations into the “Text Category Factor” are worth pursuing.

In the following chapter, I will explore a third issue, which I refer to as the “Language Factor,” to determine the extent to which this has an impact on fuzzy matching in TMs.

5 The Language Factor

Another issue that may have an impact on fuzzy matching is the question of the language of translation, and this will be the subject under investigation in this chapter. As noted in section 2.3, languages have different characteristics, such as different stylistic or structural preferences. As a result, some languages may be more amenable than others when it comes to being processed with the help of a TM system. I will examine the “Language Factor” by comparing the results of fuzzy matches generated from a French-to-English TM database against those generated from an English-to-French TM database. The methodology and results of this small experiment will be presented in the following sections, as will some preliminary guidelines that TM users may need to consider when working with TM systems in different languages.

5.1 Language Selection

Ideally, for this experiment, it would have been nice to be able to test a range of different languages; however, I have limited this investigation to the languages of English and French for practical reasons. Firstly, these two languages correspond to my own principal working languages. In addition, to carry out this experiment, it will be necessary to obtain a reasonable amount of previously translated material of a suitable nature (see Chapter 3) which can be loaded into a TM database. I know that I can access this type of material in both English and French without much

difficulty; therefore, this is a pragmatic reason for choosing to work with these languages.

This choice of languages will permit me to make a general observation about whether or not different languages have a different impact on fuzzy matching. More specifically, it will allow me to test the hypothesis that for the same type of text and TM database, translating from French to English will require the use of a lower fuzzy match threshold than will translating from English into French.

5.2 CAT Tool Selection

The TM system selected for use in the “Language Factor” experiment is Trados for the simple reason that, in the “System Factor” experiment described in chapter 3, Trados was the system that generated the greatest number of fuzzy matches.

5.3 TM Database and Source Text Selection

A number of different TM databases had already been created for the experiments described in Chapters 3 and 4, including a database of appeal decisions, a database of job advertisements, and a database of the collective agreement of the Association of Professors of the University of Ottawa. Of these, the one that generated the greatest number of fuzzy matches was the appeal decisions database, so I decided to reuse this resource for the “Language Factor” experiment.

However, since the original appeal decisions database was created using French as the source language and English as the target language, it was necessary to create

a second TM database where English was the source and French the target. This was necessary because, although TM systems can be used with virtually any language combination, the material in the TM database is always indexed according to a specified source language. This means that all searches can only be carried out in that indexed language, although the results of the search are displayed in a bilingual format (see section 1.2.2.5).

In order to be able to search in the other direction, a separate database must be created, which can contain exactly the same texts, but in which the indexing is done using the other language. So in the case of the original appeal decision database, the indexing had been carried out on the French-language texts, since French was specified as the source language. I therefore took the same pairs of texts but created a new TM database in which I specified English as the source language, and which was then subsequently indexed using the English-language material. This new English-French appeal decisions database could now be searched using patterns taken from the English texts. These two databases were the basis for the comparative “Language Factor” experiment.

As explained in section 3.2.1, the texts that were used in the TM database compilation came from the website of Citizenship and Immigration Canada (CIC). Corresponding English and French versions of texts were downloaded from the *Reflex* database for inclusion in the appeal decisions TM database. In the “System Factor” experiment outlined in Chapter 3, a number of additional French language texts were selected from the database to be used as new source texts to be

translated from French into English. For this “Language Factor” experiment, it was necessary to identify the corresponding English language versions of those texts so that these could be used as the new source texts to be translated from English into French.

5.4 Testing Procedure

Once the source texts had been gathered, the Trados fuzzy match setting was set to the lowest possible value (i.e., 30%) in order to maximise the number of fuzzy matches retrieved. Fuzzy matches were generated in both language directions using the appropriate source texts and TM databases. The findings were then recorded for later analysis, and the detailed results for each of the language directions can be viewed in Appendix 5.1.

5.5 Data Analysis

This section will present an analysis of the data generated from the “Language Factor” experiment. It will be analyzed from both a quantitative and a qualitative viewpoint.

5.5.1 Quantitative Analysis

Table 5.1 summarises the total number of fuzzy matches retrieved for the corresponding 9 segments taken from the English source texts and the French source texts.

	French-to-English	English-to-French
Segment 1	4	2
Segment 2	2	4
Segment 3	3	3
Segment 4	3	2
Segment 5	3	1
Segment 6	2	2
Segment 7	3	0
Segment 8	4	5
Segment 9	5	5
TOTAL	29	24

Table 5.1: Number of fuzzy matches produced for each segment in each language direction

These findings show that, when translating from French into English, a greater number of fuzzy matches (29) were retrieved as compared to translating from English into French (24). For 4 out of the 9 segments, the French-English language direction produced a greater number of fuzzy matches, while in 3 cases, the same number of matches were produced for both language combinations, and in only 2 out of 9 instances were a greater number of fuzzy matches generated for the English-French direction.

Moreover, there was one segment (#7) for which no fuzzy match was retrieved whatsoever when working from English into French. Therefore, we could say that, overall, a French-English translator is presented with a greater amount of information from TM fuzzy matches than is an English-French translator.

However, as noted in section 1.2.2.5, fuzzy matches are not all equal. Some can be very noisy and therefore provide little assistance to a translator. Table 5.2 presents

a summary of the findings with regard to the percentage scores of the “best” and “second-best” fuzzy matches retrieved for each segment.

	French-to-English		English-to-French	
	Highest %	2 nd highest %	Highest %	2 nd highest %
Segment 1	82%	64%	70%	57%
Segment 2	44%	41%	59%	58%
Segment 3	87%	80%	84%	79%
Segment 4	74%	64%	66%	55%
Segment 5	45%	35%	33%	No match
Segment 6	59%	45%	64%	45%
Segment 7	46%	40%	No match	No match
Segment 8	41%	38%	46%	40%
Segment 9	67%	64%	71%	67%
AVERAGE	61%	52%	55%	45%

Table 5.2 – Fuzzy match percentage values per segment

These findings reveal that in addition to generating a greater number of fuzzy matches, the French-to-English language direction also produced higher percentage matches on average. The mean score for fuzzy matches generated for the French-English combination was 61% for the “best” matches and 52% for the “second-best” matches, while for the English-French combination, the scores were 55% for the “best” matches and 45% for the “second-best” matches. In addition, in 5 out of 9 cases, the French-English direction produced higher scores for both the “best” matches and the “second-best” matches.

These findings suggest that, in addition to retrieving a greater amount of information when working from French into English, translators working in this direction will also be presented with a higher quality of information than their English-to-French counterparts. This would seem to disprove my initial hypothesis that translators working from French into English would need to set the fuzzy match

value to a lower setting. Rather, this data suggests the opposite would be true; in other words, it is the English-to-French translators who would need to set their fuzzy match value to a lower percentage.

5.5.2 Qualitative Analysis

To determine why the French-to-English matches are generating higher scores than the English-to-French matches, it is necessary to look at the actual fuzzy matches that were proposed in both language directions. In 5 out of 9 cases, for the “best match,” the very same translation unit (TU) (i.e., the same pair of segments) was retrieved in both language directions, even though this same match produced a different score in each direction. Differences in the percentage scores can be explained in part by the higher degree of variation – both at the level of lexis and of characters – that was present in the English language segments. For example, consider the following.

- **Segment 1:** The French source text and the French segment stored in the database both included the term “*décision*.” In contrast, in the English source text, the term “*decision*” was used, while the term “*determination*” appeared in the stored TM segment. This resulted in fewer matching characters between the English-language segments and therefore a lower percentage score.
- **Segments 1 and 4:** In both of these segments, the expression “*n’avait pas respecté l’obligation de résidence*” was repeated in both the French source text and the stored TM segment, whereas various synonymous expressions

were used in English. In segment 1, the English source text used the expression “*had failed to fulfill the residency obligation,*” while the stored TM segment said “*had not complied with the residency obligation.*” Similarly in segment 4, the English source text stated “*had failed to satisfy the residency obligation,*” while the expression “*had not complied with the residency obligation*” appeared in the stored segment. This use of synonymous expressions resulted in fewer matching characters overall, and therefore generated a lower score for the English segments.

- **Segment 5:** The French source text and the French segment stored in the TM database both used the long form of the term “*Règlement sur l’immigration et la protection des réfugiés.*” In contrast, the English source text used the long form “*Immigration and Refugee Protection Regulations,*” while the stored segment in the TM database contained only the abbreviated form “*IRP Regulations*” thus reducing the number of characters that matched.
- **Segment 3:** In both French and English, the fuzzy matches retrieved for the source segments are very close matches with only one term being different between the source segment and the stored TM segment. In French, the term appearing in the source text is “*conjoint de fait*” and in the stored TM segment it is “*partenaire conjugal.*” In English, the term in the source text is “*common-law partner,*” while that contained in the stored TM segment is “*conjugal partner.*” Keeping in mind that, as explained in section 1.2.2.5, a TM system operates primarily on the basis of pattern matching, we can see that there is a closer superficial match between the French terms “*conjoint*” and “*conjugal*” (i.e., both contain the string “*conj*”) than there is between the

English terms “common-law” and “conjugal” (i.e., the only common string is “co”). Therefore, the French language match generated a slightly higher score.

In 3 other instances (i.e., for segments 2, 6 and 8) different TUs were retrieved, as shown in Table 5.3. What is interesting here is that, in all 3 cases when different matches were retrieved, the percentage score was higher for the English-to-French language direction.

Segment	French-to-English database	English-to-French database
2	New ST: La demande parrainée de résidence permanente de la femme de l'appellant a été refusée parce que le mariage était une relation de mauvaise foi. FR: La demande parrainée de résidence permanente des nièces orphelines de l'appellant a été refusée parce que les demandeurs étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés. EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased. Matching value: 44%	New ST: The sponsored application for permanent residence of the appellant's wife was refused on the basis that the marriage was a bad faith relationship. EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child." FR: La demande parrainée de résidence permanente du fils de l'appellant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ». Matching value: 59%
6	New ST: Lors de l'appel, l'appellant n'a pas contesté la validité de cette mesure. FR: L'appellant n'a pas contesté la validité juridique du refus. EN: The appellant did not challenge	New ST: He did not challenge the validity of the order on appeal. EN: He did not challenge the legal validity of the decision. FR: Il n'a pas contesté la validité juridique de la décision.

Segment	French-to-English database	English-to-French database
	the legal validity of the refusal. Matching value: 59%	Matching value: 64%
8	New ST: La demande parrainée de résidence permanente présentée par les parents de l'appellant a été refusée au motif que l'appellant ne pourrait respecter son engagement d'aide et que les requérants ne pourraient subvenir eux-mêmes à leurs besoins. FR: La demande parrainée de résidence permanente de la mère de l'appellant, qui était veuve, a été refusée parce que la demandeuse était interdite de territoire pour motifs sanitaires, en ce sens qu'elle était atteinte d'arthrose dans les deux genoux. EN: The sponsored application for permanent residence of the appellant's widowed mother was refused because the applicant was medically inadmissible in that she suffered from osteoarthritis in both knees. Matching value: 41%	New ST: The sponsored application for permanent residence of the appellant's parents was refused on the ground that the appellant would be unable to fulfil his undertaking of assistance and the applicants would be unable to support themselves. EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased. FR: La demande parrainée de résidence permanente des nièces orphelines de l'appellant a été refusée parce que les demandeuses étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés. Matching value: 46%

Table 5.3 – Higher English-to-French percentage scores for the same segment

Finally, in the case of segment 7, a fuzzy match was retrieved for this segment in the French-to-English direction, but no match was found when working from English into French, as illustrated in Table 5.4. A possible explanation for this could be as follows. The match made between the two French segments has been made primarily on the basis of the term “*Loi sur l’immigration et la protection des réfugiés (LIPR)*,” which contains 50 characters.

In the English segment, there is no abbreviation included as part of the term, and the term “*Immigration and Refugee Protection Act*” itself is much shorter than its French counterpart due to the preference of English for pre-modification as compared to the use of prepositional phrases that is more common in French. Therefore, the term in English contains only 34 characters. Since the French term is almost one-third longer than the English term, it corresponded to a larger portion of the segment stored in the database (i.e., the French term made up more than 30% of the segment stored in the TM), whereas the shorter English term corresponded to barely 25% of the English segment stored in the TM database, which was below the minimum fuzzy matching value threshold of 30%, and hence it was not retrieved from the database.

Segment	French-to-English database	English-to-French database
7	New ST: Le ministre a présenté une demande d’interdiction de divulgation de renseignements aux termes de l’article 86 de la Loi sur l’immigration et la protection des réfugiés (LIPR). FR: L’appelant en a appelé d’une décision selon laquelle il n’avait pas respecté l’obligation de résidence prévue à l’article 28 de la Loi sur l’immigration et la protection des réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 46%	New ST: The Minister brought an application for non-disclosure of information under section 86 of the Immigration and Refugee Protection Act. NO FUZZY MATCHES RETRIEVED

Table 5.4 – No fuzzy match retrieved from English into French

5.6 Conclusions

After examining the findings of this investigation into both the French-to-English and the English-to-French databases, I have reached the following conclusions.

Firstly, in my initial hypothesis, I had noted that the literature (e.g. Vinay and Darbelnet 1958: 252) seems to suggest that English is a language that prefers repetition, whereas French is a language that will more often opt for variation. Accordingly, I surmised that translators working from English-to-French would be more likely to get a greater number of high-scoring matches, and could therefore afford to set the fuzzy match value to a higher setting.

In this particular set of texts, the English language texts show a greater degree of variation than the French texts; therefore, the English-to-French translators would need to set their fuzzy match value to a lower setting than would translators working from French into English. I should like to point out, however, that while my precise hypothesis was disproved by this data, the general underlying principle behind it was validated: a language that has a high degree of consistency/low degree of variation can utilise a higher fuzzy match value than a language with a low degree of consistency/high degree of variation.

In this “Language Factor” experiment, the French texts showed a lesser degree of variation, and correspondingly, it was the French-to-English segments that generated both a greater number and a higher quality of matches. Therefore, as a general rule, translators should be aware that if the source language with which

they are working is one that is prone to variation, they need to reduce the level of the minimum fuzzy match value. In contrast, if the source language is highly consistent, then they can comfortably raise the value of the minimum fuzzy match setting.

The other issue that I raised in my initial hypothesis (section 2.3) was the question of the “lengthiness” of a language. I had noted that the literature (e.g. Vinay and Darbelnet 1958: 103, 109) suggests that, because of its structural preferences, French typically requires a larger number of words than does English in order to express the same idea. I initially felt that this added length would present a disadvantage when working with TMs, feeling that a shorter segment might offer a greater chance for achieving a match, since the shorter the segment, the less room there is for variation.

However, upon further reflection fuelled by the analysis of the data from the “Language Factor” experiment, I have now come to feel that this question of length could possibly be seen as an asset for achieving a fuzzy match, rather than as a hindrance. A longer segment is less likely to present an exact match since the longer it becomes, the greater the chances that some part of the segment will differ from another segment; however, it may also simultaneously increase the chance that some part of the segment will be similar to some part of another segment.

So while longer matches are less likely to generate exact matches, they could very well be more likely to generate partial or fuzzy matches. Therefore, when it comes

to fuzzy matching, the lengthier languages may actually be more likely to generate matches, and therefore a translator working with a “lengthy” language can set the minimum fuzzy match value to a higher setting than can a translator working with a more compact language.

In the following chapter, I will explore one final issue, which I refer to as the “Degree of Fuzziness Factor,” to determine the extent to which this has an impact on fuzzy matching in TMs.

6 The Degree of Fuzziness Factor

In the previous three chapters, I have investigated whether or not the choice of system, the nature of the text, and the language of translation seem to influence the way in which TM systems generate fuzzy matches. In this chapter, I will focus on a fourth and final issue, which I refer to as the “Degree of Fuzziness Factor.” The approach taken in this chapter will depart somewhat from that taken in the three preceding chapters. In this case, I set out to prove or disprove an axiom²² that is currently present in the industry. This axiom corresponds to the belief that a higher percentage match is a “better” match in terms of its usefulness to translators. The following sections present the methodology used in this investigation, as well as an analysis of the results. Based on these findings, I will be able to determine whether what is currently an axiom can actually be considered as fact.

6.1 Participant Selection

The basic methodology used this experiment will be to provide translators with a selection of fuzzy matches that have been retrieved for a given segment and to ask them which of these matches they think is the most useful for helping them to translate that segment.

When selecting participants to take part in this experiment, I wanted to find a group of translators who worked from French to English, who were familiar with TM tools

²² As discussed in section 2.4, an axiom is a statement that is accepted as a truth but which has not been empirically proved to be true.

and the concept of fuzzy matching, and who were relatively homogeneous in terms of their translator training and experience.

I was able to identify 10 volunteer participants, all of whom are Anglophones and all of whom graduated with an MA in Translation from the University of Ottawa between 2002 and 2006. Furthermore, all of these participants had followed a course in translation technology as part of their studies, and they were therefore all familiar with TM systems and fuzzy matching.

6.2 CAT Tool Selection

Once again, the TM system that was selected for this experiment is Trados since, during the system comparison carried out in Chapter 3, Trados generated a greater number of instances where multiple fuzzy matches were retrieved for a given segment.

6.3 TM Database Selection

The two TM databases first introduced in Chapter 3 (sections 3.2.1 and 3.2.2), which were initially created and used for the “System Factor” experiments, were used once again for this investigation into the “Degree of Fuzziness Factor.” These two databases were the appeal decisions database and the job advertisements database.

6.4 Testing Procedure

As in the previous experiments, Trados's minimum fuzzy match value was set to 30% in order to generate the maximum number of fuzzy matches. A considerable number of matches were generated using both the appeal decisions database and the job advertisements database. Because the participants in this experiment were acting in a volunteer capacity, I selected only 10 sets of matches²³ to include in the survey, so as not to overburden the volunteers. I chose to use the sets of matches that generated the highest percentages. Accordingly, 7 sets of multiple fuzzy matches were taken from appeal decisions database, and the remaining 3 sets of multiple fuzzy matches came from the database of job advertisements.

A "test" containing 10 questions was devised, and each question had the same format. Essentially, participants were given a French segment that needed to be translated into English. They were also shown multiple fuzzy matches for that segment that had been retrieved using a TM system; however, the participants were not given the percentage of each match. The participants were asked to examine all the proposed fuzzy matches and to rank these according to which they felt would be the most useful for helping them to translate the French-language segment (where a ranking of 1 = most helpful, 2 = 2nd most helpful, etc.). An example of the question format is given below, while the entire "test," along with the specific instructions given to participants, can be viewed in Appendix 6.1.

²³ A "set of matches" could include either two or three different fuzzy matches that were retrieved for a given segment. In this experiment, 5 of the 10 match sets consisted of two fuzzy matches, and the other 5 consisted of 3 fuzzy matches.

EXAMPLE:

New source sentence to be translated:

Vous vous intéressez à l'économie et à la finance et comptez au moins cinq années d'expérience comme traducteur dans ces domaines ou d'autres domaines pertinents?

Proposals from the translation memory system:

FR: La préférence sera accordée aux personnes qui ont au moins deux années d'expérience encadrée en traduction, ou l'équivalent en formation et en expérience.

[enter your ranking]

EN: Preference will be given to applicants with at least two years of supervised experience in translation, or equivalent in training and experience.

FR: Vous avez au moins deux années d'expérience comme traducteur dans un ou plusieurs des domaines suivants :

[enter your ranking]

EN: As the ideal candidate, you are a translator with a minimum of two years experience in at least one of the following fields:

Figure 6.1 – Example of the question format given to translators

A space was provided under each new source text segment where the translators could make rough notes or write a draft translation if they found that exercise to be useful for assisting them to evaluate the helpfulness of the various TM proposals. There was also a space following each question where the participants could make any comments they deemed relevant.

An important part of this experiment was drafting clear instructions for the participants. It was emphasised to the translators that they were *not* being asked to rank which of the proposed fuzzy matches was the most accurate translation of the new source sentence. Rather, they were to rank each proposed translation according to which they found to be the most helpful (and 2nd most helpful, etc.) when it came to translating the new source segment from French into English. The “test” was sent to participants via email and they were given seven days to return their responses.

6.5 Data Analysis

Responses were received from all 10 participants; however, one of the participants did not follow the instructions and did not provide a clear ranking of the proposed fuzzy matches, so the data from that participant was not retained. Therefore, the following sections will analyze only the data from the other 9 participants.

6.5.1 Analysis according to participants

Table 6.1 summarises the rankings provided by those 9 participants.

Participant #	Number of times highest % match was selected	Number of times 2 nd highest % match was selected	Number of times 3 rd highest % match was selected
1	7	3	0
2	4	6	0
3	6	4	0
4	8	2	0
5	5	5	0
6	6	4	0
7	8	2	0
8	6	4	0

9	6	4	0
TOTAL (/90 segments)	56/90	34/90	0/90
% of segments selected	62.2 %	37.8%	0%

Table 6.1 – Results of Multiple Fuzzy Match Experiment

These results indicate that 7 out of the 9 participants selected the highest percentage match as being the most helpful match the majority of the time. One participant (participant 2) selected the second highest percentage match as being the most helpful match the majority of the time. The remaining participant (participant 5) selected the highest percentage match half of the time, and the second highest percentage match the other half of the time.

It is interesting to note that of the 7 participants who felt that the highest percentage match was most often the most helpful match, only 3 of the 7 chose the highest percentage match more than 70% of the time, while the other 4 participants all selected the highest percentage match as the most helpful match in only 60% of the cases.

Looking at the global results, we can see that the highest percentage match was selected by participants as being the most helpful match in 62.2% percent of the cases, while the remaining 37.8% of the time, the 2nd highest match was considered to be the most useful. Furthermore, although 5 of the 10 match sets contained 3 fuzzy matches, there were no cases where a participant selected the 3rd highest match as being the most useful match.

On the basis of this data, it would seem that although the majority of participants did find the highest percentage match to be the most helpful match on the greatest number of occasions, it was not by any means rare for a participant to select the 2nd highest percentage match to be the most useful for translating a given segment. In fact, close to 40% of the time, the 2nd highest match proved to be the most helpful match as judged by the translators.

6.5.2 Analysis according to segments

Another way of looking at the data is according to the number of times that the highest percentage match was/was not selected as the most helpful match for a given segment. This data is summarised in Table 6.2.

Segment #	Number of Participants who DID choose the highest % match for the segment	Number of Participants who did NOT choose the highest % match for the segment	% of time that the highest % match was selected as the most helpful match for the segment
1	7	2	77.8%
2	6	3	66.7%
3	5	4	55.6%
4	8	1	88.9%
5	7	2	77.8%
6	8	1	88.9%
7	0	9	0%
8	2	7	22.2%
9	6	3	66.7%
10	7	2	77.8%

Table 6.2 – Number of Participants Who Did Not Choose the Highest % Match

The information in Table 6.2 reveals a number of interesting points. Firstly, we can see that for every single one of the 10 segments, there was at least one participant who did not choose the highest percentage match as the most helpful match.

Secondly, there was one segment (#7) where all the participants felt that the 2nd highest percentage match was actually the most useful and another segment where 7 of the 9 participants identified the 2nd highest match as being more helpful. However, for 8 out of the 10 segments, the majority of the participants did in fact identify the highest percentage match as being the most useful match.

6.5.3 Analysis of participants' comments

As indicated, above, participants were invited to make any comments that they found to be relevant with regard to their experience ranking the matches. Not every participant provided a comment for every question, but a number of points are worth noting based on the comments that were received. The full details of the comments can be found in Appendix 6.2.

For segments 4, 6, 8 and 10, a number of participants commented that, although they had provided a ranking as indicated in the instructions, they actually did not find any of the fuzzy matches to be particularly helpful. It is therefore not surprising to note that all of these segments had relatively low fuzzy matching scores, ranging from 37-47%. This would seem to confirm that a low fuzzy matching score is limited in the amount of useful information that it can provide to the translator.

Another interesting observation is that for segments 2, 3, 5 and 7, several translators noted that the top two matches were nearly equal in terms of the help that they provided. A look at the scores for these segments shows that the

percentage of difference between the highest and second-highest fuzzy matches in these segments was quite small, ranging from 3% to 9%.

Other points of note that came out in the comments are that translators preferred segments that required less editing, and that while translators were always interested in matches that provided useful terminology, they also appreciated receiving help with issues such as style and sentence structure.

6.6 Conclusions

The principal objective of this chapter was to determine whether or not there is any empirical evidence to support the axiom purporting that in cases where there are multiple fuzzy matches for a given segment, the one with the highest percentage match value will be the one that is most useful for helping a translator to formulate an appropriate translation. Based on the data collected from the experiment presented in this chapter, it would seem that this axiom is in fact generally true. The findings of the experiment revealed that approximately 62% of the time, translators found the highest percentage match to be the most useful. However, there is also ample evidence to suggest that in many cases, the translator may benefit from looking beyond the highest percentage match and consulting the 2nd highest percentage match since nearly 40% of the time, translators found that this match was the one that gave them the most assistance with the translation task, and for 2 of the 10 segments used in this experiment, nearly all of the translators found the 2nd highest percentage match to be the most helpful.

Another point that could be worth considering is the fact that for every single one of the 10 segments used in this experiment, there was at least one translator (and in the vast majority of cases, there was more than one) who found the 2nd highest match to be more useful than the highest percentage match. Therefore, it could be advantageous for all translators to consider consulting more than one fuzzy match in cases where multiple matches are proposed. It would seem that any given translator will, at least some of the time, find the information provided in the 2nd highest match to be more useful than the information provided in the highest scoring match.

Moreover, it is useful to note that translators in this experiment never once selected the 3rd highest percentage match, which could be an indication that even in cases where translators are offered more than two fuzzy matches for consultation; they need not waste their time looking beyond the second match.

Finally, although the goal of this experiment was primarily to determine whether or not there was any truth to the axiom that claims that the highest scoring fuzzy match is always the most useful, there is another conclusion that can be drawn. Because translators frequently find that the 2nd highest match is helpful, they might want to take care not to set the match value too high in order to ensure that the 2nd highest matches do not fall below the threshold.

This concludes the practical experimentation section of this thesis. In the final chapter, I will provide an overall evaluation of this research, as well as suggestions as to how it might be improved and extended in the future.

7 Conclusion

Each of the preceding four chapters contained the description and results of an experiment conducted to learn more about a certain aspect of fuzzy matching. In this chapter, I will present an overall assessment of the work carried out as part of this research project. In addition, I will offer some general concluding remarks and make some suggestions for ways in which this work could be expanded in the future.

7.1 Overall Assessment

The general objective of this thesis as stated in section 0.1 was to gain a deeper understanding of fuzzy matching in the context of TMs. To achieve this goal, I set out to investigate four specific factors to determine the extent to which they had an influence on selecting the ideal minimum match value for fuzzy matching: the “System Factor,” the “Text Category Factor,” the “Language Factor” and the “Degree of Fuzziness Factor.” I feel that this goal has definitely been met: the four experiments allowed me to gain a much greater insight into fuzzy matching and the factors that affect it. The upcoming sections will summarise the key points that were learned.

A second goal of the thesis was to propose some preliminary guidelines that could help translators to decide where to set the fuzzy match threshold when using a TM system. I was less successful at meeting this goal because, as will be discussed

below in section 7.1.5, the findings of one experiment were sometimes at odds with the findings of another.

7.1.1 Chapter 3 – The System Factor

A test of two TM systems was carried out in Chapter 3, using two separate TM databases. The two investigations produced similar results, which means that I can generalise the conclusions to a certain degree. First of all, it appears that of the two TM systems, Trados seems to offer the translator more in terms of the number of fuzzy matches presented to the user. These results included fuzzy matches where Fusion Translate did not present any, as well as multiple options in Trados terms of fuzzy matches for the same given segment. Also, when the two systems both generated a match, the matches proposed by Trados were generally of a higher quality than those proposed by Fusion.

Despite the fact that the results presented in Chapter 3 are extremely limited (in that only two systems were examined, and only two databases were used), this thesis has shown that the TM system that a translator either chooses or is asked to use definitely has an impact when it comes to the number and quality of fuzzy matches that are generated. In other words, the “System Factor” is definitely a factor that must be considered when selecting a fuzzy match threshold.

7.1.2 Chapter 4 – The Textual Characteristics Factor

Chapter 4 presented the results of an experiment carried out on two different categories of texts – “Recycled Texts” and “Revised Texts.” This particular

experiment demonstrated that the fuzzy matches generated produced two different types of results, each helpful in their own right. The “Recycled Texts” generated a much larger number of fuzzy matches that were generated by the “Revised Texts.” These results may be considered by some to be helpful, since the translator is likely to have a greater number of matches at his/her disposal in order to help with the translation task. However, few of the fuzzy matches generated in the “Recycled Texts” produced matches of a relatively high quality. In contrast, the matches retrieved using the “Revised Texts” produced fuzzy matches of a relative high quality, albeit there were considerably fewer matches as compared to the “Recycled Texts.”

Based on the results of this small experiment, it appears that translators working with “Recycled Texts” would benefit from using a lower fuzzy match threshold than would translators using a TM system working with “Revised Texts.” Therefore, it appears that the “Category of Text Factor” is indeed something that a translator needs to take into account when determining the minimum fuzzy match setting.

7.1.3 Chapter 5 –The Language Factor

The fifth chapter of this thesis contained a comparison of fuzzy matches generated from TM databases containing different language directions. The two databases contained exactly the same texts, but the first had French as a source language and English as a target language, while the second was the reverse. It was found that for this set of texts, translators working from English to French would likely be able to retrieve a greater number of high-scoring matches than would their French-to-English counterparts, and the former could therefore afford to set the fuzzy match

value to a higher setting. This experiment was very limited in terms of the languages involved, and it also contained only one type of text/database. Therefore, rather than attempting to make a generalised statement about French and English, it seems wiser to draw a more abstract conclusion: if the source language is one that is prone to variation (lexical or syntactic), a lower minimum fuzzy match value would be advantageous for generating helpful fuzzy matches. On the contrary, if the source language is highly consistent, then the value of the minimum fuzzy match setting can be raised.

A second conclusion that was reached based on the findings of the experiment conducted in chapter 5 concerns sentence length. Lengthier²⁴ languages may actually be more likely to generate fuzzy matches, and a higher minimum fuzzy match value could therefore benefit a translator working with a “lengthy” language, as opposed to a translator working with a more “compact” language.

The conclusion can therefore be drawn that the “Language Factor” does indeed come into play when setting the fuzzy match threshold in a TM system.

7.1.4 Chapter 6 – The Degree of Fuzziness Factor

The overall objective of chapter 6 was to prove or disprove the axiom which claims that when multiple fuzzy matches are generated for a given segment, the highest scoring fuzzy match is always the most helpful to the TM user. The results showed that this axiom is generally true (62% of the time), but not overwhelmingly so. The

²⁴ For example, languages that use prepositional phrases as opposed to pre-modification.

data revealed that on many occasions, a translator can benefit from examining the data contained in the 2nd highest percentage fuzzy match. Therefore, I consider that the "Degree of Fuzziness Factor" also comes into play when setting the fuzzy match threshold in a TM system.

7.1.5 Challenges Associated with the Development of Guidelines for Translators to Help Them Set the Fuzzy Match Threshold

It has been noted previously that fuzzy matching is a complex and poorly understood process. Although the work presented in this thesis has gone some way toward providing a deeper understanding of this phenomenon, I cannot claim to have demystified it completely. My original hope was that I would be able to develop a set of straightforward guidelines that translators could use to help them determine the ideal fuzzy match threshold to use in a variety of circumstances. Based on empirical investigations, I was successfully able to determine, for example, that a translator working on a "Revised Text" could use a higher fuzzy match threshold than a translator working on a "Recycled Text," and that a translator working with a "consistent" language could use a higher fuzzy match threshold than a translator working on a "variable" language.

It was relatively straightforward to come up with guidelines for each of the individual factors, but things quickly became complicated when I tried to consider the various factors together. For example, what should a translator do if he/she is working on a "Recycled Text" (calling for a low fuzzy match threshold) BUT with a source language that is consistent (thus allowing for a higher fuzzy match

threshold)? Because a translation situation includes all of the factors, there may often be cases where the guidelines that apply to one of those factors cancel out the guidelines that apply to another. Clearly, as discussed in the next section, more work remains to be done.

7.2 Suggestions for Future Research

There are several ways in which the research reported in this thesis can be improved and/or expanded.

7.2.1 A Holistic Analysis of the Factors

As noted in the previous section, a study of each of the individual factors in isolation represents a good starting point for learning more about how these factors affect fuzzy matching, but in a next step, it will be necessary to undertake a more holistic analysis of the factors, studying how they interact, and in cases where there is an apparent conflict, determining which factor should take precedence. Such a study is beyond the scope of this MA thesis, but it would be a very interesting and valuable contribution for a future project.

7.2.2 Scaling Up the Investigations

A second major area in which this research could be expanded entails scaling up all parts of the project. I feel that the methodologies used in each of the four experiments were sound, but the size of these studies was too small to permit me to draw any truly generalizable or definitive conclusions.

7.2.2.1 Number of TM Systems Tested

A similar type of experiment on TM systems such as the one carried out in chapter 3 should be carried out using a wider range of commercially-available TM systems, such as the ones mentioned in section 1.1.

7.2.3 Number of Categories

The results in chapter 4 were obtained using two categories of textual characteristics. A wider-ranging experiment might also investigate all four categories as described in section 4.1.

7.2.4 Number of Languages

A future project carried out in the style of chapter 5 would benefit from a wider-ranging assortment of languages. Of course, such a project would require additional resources and translators able to work in a variety of language combinations. As noted in section 0.1, efforts by companies to increase their global sales often necessitate the translation of documentation from English into several languages, many more than just the two official languages of Canada. Therefore, such a project should take this into consideration.

7.2.5 Larger Participant Pool

Further investigations requiring the input of external translators such as the one carried out in chapter 6 would most definitely benefit from using a much larger pool of participants. The test would also benefit from being administered in a controlled setting (i.e., a setting with time limits, a Question-&-Answer session in order to clear

up any confusion, etc.). Different experiments addressing different types of translators would also be interesting (e.g. to see if novice translators have different needs than more experienced translators, or if L1 translators have different needs than L2 translators).

7.2.6 Size of TM Database

Compiling TM databases is a time-consuming and labour-intensive job. I feel that the TM databases used in this experiment were appropriate for an MA thesis, but in order to obtain results that are more definitive and conclusive, it would be necessary to build TM databases that are significantly larger in size.

7.3 Concluding Remarks

Each year, more and more translators and those who employ them are beginning to recognise the enormous impact that technology is having on this field, and to learn that, when used correctly, technology can help to speed up and facilitate this task. Tools such as the TM systems that have been the focus of this study have the potential to revolutionise the translation industry. Further knowledge, empirical investigation, and mastery of this technology can only help take translation and translation technologies to a higher level of acceptance and respect in academic and professional translation circles worldwide.

References

- Allen, Jeffrey. (2003) "Post-editing." In: Harold Somers, ed. *Computers and Translation: A Translator's Guide*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 297-317.
- Arthern, P. J. (1981) *Aids Unlimited: The Scope for Machine Aids in a Large Organization*. Aslib Proceedings, vol. 33, no. 8, 309-319.
- Austermühl, Frank. (2001) *Electronic Tools for Translators*. Manchester: St. Jerome Publishing.
- Bédard, Claude. (2000) "Mémoire de traduction cherche traducteur de phrases." *Traduire*, 186: 41-49. [http://www.terminotix.com/t_fr/index.htm]
- Benis, Michael. (1999) *Translation Memory from O to R*. "TransRef – The Translation Reference Centre." [<http://transref.org/default.asp?docsrc=/u-articles/Benis3.asp>]
- Benis, Michael. (2000) "Extended comparative review of translation memory products- part 2" [<http://www.transref.org/default.asp?docsrc=/u-articles/Benis4.asp>]
- Bowker, Lynne. (2002a) *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press.
- Bowker, Lynne. (2002b) "Information Retrieval in Translation Memory Systems: Assessment of Current Limitations and Possibilities for Future Development." *Knowledge Organization*, 29 (3/4), 198-203.
- Bowker, Lynne. (2005) "Productivity vs Quality? A Pilot Study on the Impact of Translation Memory Systems," *Localization Focus* 4(1), 13-20.
- Bowker, Lynne and Barlow, Michael. (2004) "Bilingual concordancers and translation memories: A comparative evaluation" In *Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training*, 52-61. [http://www.ifi.unizh.ch/cl/yuste/lr4trans-2/WKS_PAPERS/8.pdf]
- Brockmann, Daniel. (1999) "Translation Memories as True Databases: Present and Future." *The European Language Resources Association Newsletter* 4.3: July – September 1999, 9-11.
- Canadian Translation Industry Sectoral Committee (CTISC). 1999. "Survey of the Canadian Translation Industry." URL: <http://www.uottawa.ca/associations/csict/princi-e.htm>. [Accessed October 23, 2006]

Chu, Heting. (2003) *Information Representation and Retrieval in the Digital Age*. Medford, New Jersey: Information Today, Inc.

Circé, Karine. (2005) *Traduction automatique, mémoire de traduction ou traduction humaine? Proposition d'une approche pour déterminer la meilleure méthode à adopter, selon le texte*. MA Thesis, University of Ottawa.

Cronin, Michael. (2003) *Translation and Globalization*. London and New York: Routledge.

De Vries, Arjen-Sjoerd. (2002) "Getting Full or Fuzzy? The payment issue with full matches generated by translation memory systems." *Language International* (June 2002), 44-47.

Falcone, Suzanne. (1998) "Translation Aid Software: Four Translation Memory Programs Reviewed" *Translation Journal* 1(2)
[<http://accurapid.com/journal/03TM2.htm>]

Esselink, Bert. (2000) *A Practical Guide to Localization*. Rev. Ed. Amsterdam and Philadelphia: John Benjamins.

Garcia Izquierdo, Isabel. (2000) "The Concept of Text Type and Its relevance to Translator Training." *Target*, vol. 12, no. 2, 283-295.

Gow, Francie. (2003) *Metrics for Evaluating Translation Memory Software*. MA thesis, University of Ottawa. [<http://www.chandos.ca/thesis.html>]

Höge, Monika. (2002) "Towards a Framework for the Evaluation for Translators' Aids Systems," Ph.D Dissertation. University of Helsinki
[<http://ethesis.helsinki.fi/julkaisut/hum/yleis/vk/hoge/>]

Kay, Martin. (1997[1980]) "The Proper Place of Men and Machines in Language Translation." *Machine Translation*. Vol. 12(1/2), 3-23.

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

Macklovitch, Elliot and Russell, Graham. (2000) "What's Been Forgotten in Translation Memory" in *Envisioning Machine Translation in the Information Future*. 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000. J.S. White (Ed.) Berlin: Springer Verlag, 137-146.

Macklovitch, Elliot and Russell, Graham. (2002) "Calculating Return on Investment in Tools." *MultiLingual Computing and Technology*. Vol. 12, issue 3, 41-44.

Melby, Alan K., with Terry C. Warner. (1995) *The Possibility of Language: A Discussion of the Nature of Language with Implications for Human and Machine Translation*. Amsterdam and Philadelphia: John Benjamins.

Mogensen, Else. (2000) "Orwellian Linguistics: How using computer-aided translation tools impacts the target language." *Language International* (October 2000), 28-31. [http://www.erikseninc.com/pdf/Orwell_Mogensen.pdf]

O'Brien, Sharon. (1998) "Practical Experience of Computer-Aided Translation Tools in the Localization Industry." In Lynne Bowker, Michael Cronin, Dorothy Kenny, and Jennifer Pearson (Eds.), *Unity in Diversity? Current Trends in Translation Studies*, Manchester: St. Jerome Publishing, 115-22.

Quah, Chiew Kin. (2006) *Translation and Technology*. Basingstoke, UK: Palgrave Macmillan.

Rico, Celia. (2000) "Evaluation metrics for translation memories" *Language International*, vol. 12, no. 6, December 2000, 36-37.

Ross, Timothy J, Jane M. Booker, W. Jerry Parkinson. (Editors) (2002) *Fuzzy Logic and Probability Applications: Bridging the Gap*. Philadelphia, USA: Society for Industrial and Applied Mathematics; Alexandria, Virginia, USA: American Statistical Association.

Schäler, Reinhard. (2001) "Beyond Translation Memories" In Carl, M. and Way, A. (eds) *Proceedings of the Workshop on Example-Based Machine Translation*, held at the *Machine Translation Summit VIII in Santiago de Compostela, Spain*, Proceedings published on CD-ROM.

Schwarzl, Anja. (2001) *The (im)Possibilities of Machine Translation*. Frankfurt am Main, Germany; New York, USA: Peter Lang.

Somers, Harold. (2003a) "The translator's workstation," in H. Somers, ed. *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins, 13-30.

Somers, Harold. (2003b) "Translation memory systems," in H. Somers, ed. *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins, 31-47.

Trosborg, Anna. (1997) "Introduction," in Anna Trosborg, *Text Typology and Translation*. Philadelphia: John Benjamins, vii-xvi.

Trujillo, Arturo. (1999) *Translation Engines: Techniques for Machine Translation*. London, UK: Springer Publishing Company.

Wallis, Julian. (2006) *Interactive Translation vs Pre-Translation in the Context of Translation Memory Systems: Investigating the Effects of Translation Method on Translator Productivity, Quality and Translator Satisfaction*. MA Thesis, University of Ottawa.

Webb, Lynn E. (1998) "Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis" MA Thesis, Monterey Institute of International Studies, USA. [<http://www.webbsnet.com/translation/thesis.html>]

Yunker, John. (2003) *Beyond Borders: Web Localization Strategies*. Indianapolis, Indiana: New Riders.

Zerfass, Angelika. (2002a) "Evaluating Translation Memory Systems: in Proceedings of the LREC 2002 Language Resources in Translation Work and Research Workshop, ed. Elia Yuste, May 28, 2002, Las Palmas de Gran Canaria, 49-52.

Zerfass, Angelika. (2002b) "Comparing Basic Features of TM tools" Language Technology Supplement, Multilingual Computing and Technology #51, vol. 13, Issue 7, 11-14.

Zervaki, Thei. (2002) *Globalize, Localize, Translate: Tips and Resources for Success*. Bloomington, IN: 1st Books

APPENDIX 3.1 – Job Advertisements Fuzzy Matching Results (Trados vs. Fusion Translate)

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
1) Vous vous intéressez à l'économie et à la finance et comptez au moins cinq années d'expérience comme traducteur dans ces domaines ou d'autres domaines pertinents?	FR: Vous avez au moins deux années d'expérience comme traducteur dans un ou plusieurs des domaines suivants : EN: As the ideal candidate, you are a translator with a minimum of two years experience in at least one of the following fields: Matching value: 36% FR: La préférence sera accordée aux personnes qui ont au moins deux années d'expérience encadrée en traduction, ou l'équivalent en formation et en expérience. EN: Preference will be given to applicants with at least two years of supervised experience in translation, or equivalent in training and experience. Matching value: 33%	No corresponding fuzzy matching results No corresponding fuzzy matching results
2) La Banque du Canada souscrit aux principes de l'équité en matière	FR: Nous souscrivons au principe de l'équité en matière d'emploi.	No corresponding fuzzy matching results

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
d'emploi.	EN: We are committed to employment equality. Matching value: 69%	No corresponding fuzzy matching results
3) POSTE DE RÉVISEUR ANGLAIS-FRANÇAIS À TORONTO	FR: POSTE DE TRADUCTEUR ANGLAIS-FRANÇAIS À OTTAWA EN: ENGLISH TO FRENCH TRANSLATOR (OTTAWA) Matching value: 75%	FR: POSTE DE TRADUCTEUR ANGLAIS-FRANÇAIS À OTTAWA EN: ENGLISH TO FRENCH TRANSLATOR (OTTAWA) Matching value: 73%
4) Baccalauréat en traduction ou diplôme d'études universitaires dans une discipline connexe	FR: Un diplôme universitaire en traduction ou dans une discipline connexe est exigé. EN: A university degree in translation or related field is required. Matching value: 46%	FR: * Baccalauréat en traduction EN: * Bachelor Degree in translation Matching value: 100% EXACT MATCH
5) Au moins 2 années d'expérience comme réviseur en plus de 5 années en traduction supervisée	FR: Au moins 1 à 3 ans d'expérience en traduction en milieu commercial, de préférence en assurances, services financiers EN: Minimum of 1-3 years experience as a translator within a business setting,	No corresponding fuzzy matching results No corresponding fuzzy matching results

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
	preferably in insurance, financial services Matching value: 36% FR: Vous avez au moins deux années d'expérience comme traducteur dans un ou plusieurs des domaines suivants : EN: As the ideal candidate, you are a translator with a minimum of two years experience in at least one of the following fields: Matching value: 34%	
6) Avoir le souci du détail, de l'exactitude et du style	FR: La connaissance du français et de l'anglais est essentielle. EN: Knowledge of English and French is essential. Matching value: 34%	No corresponding fuzzy matching results No corresponding fuzzy matching results
7) maîtrise du français et de l'anglais, parlé et écrit;	FR: La connaissance du français et de l'anglais est essentielle. EN: Knowledge of English and French is	FR: La connaissance du français et de l'anglais est essentielle. EN: Knowledge of English and French is

New ST segment to be translated.	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
	essential. Matching value: 60% FR: Excellente connaissance de l'anglais et du français. EN: Excellent command of English and French. Matching value: 56% FR: Traducteur ou traductrice conseil, de l'anglais au français EN: Translator/Language Advisor, English to French Matching value: 38%	essential. Matching value: 55%
8) diplôme universitaire en traduction;	FR: Vaste expérience en traduction EN: Extensive translation experience Matching value: 52% FR: Études postsecondaires en traduction	No corresponding fuzzy matching results No corresponding fuzzy matching results

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
	EN: Post secondary education in translation Matching value: 52%	
9) FORD CANADA SOUSCRIT A L'EQUITÉ EN MATIÈRE D'EMPLOI	FR: Nous souscrivons au principe de l'équité en matière d'emploi. EN: We are committed to employment equality. Matching value: 43%	No corresponding fuzzy matching results No corresponding fuzzy matching results
10) POSTE RELEVANT DE :	FR: DATE DE CLÔTURE : EN: DEADLINE FOR APPLICATIONS: Matching value: 42%	No corresponding fuzzy matching results No corresponding fuzzy matching results
11) Expérience en rédaction, révision et en traduction	FR: Expérience en révision EN: Previous revision experience Matching value: 37%	No corresponding fuzzy matching results No corresponding fuzzy matching results
12) Réviseur(e) (anglais – français)	FR: Bac en traduction (français-anglais) EN: Bachelor's Degree in Translation (French to English) Matching value: 44%	No corresponding fuzzy matching results No corresponding fuzzy matching results

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
13) Toutes les candidatures seront traitées confidentiellement et dans le respect du principe d'équité en matière d'emploi.	FR: Nous souscrivons au principe de l'équité en matière d'emploi. EN: We are committed to employment equality. Matching value: 35%	No corresponding fuzzy matching results No corresponding fuzzy matching results
14) Si ce défi vous intéresse, vous pouvez acheminer votre curriculum vitae selon votre convenance à l'adresse courriel :	FR: Pour soumettre votre candidature à ce poste, veuillez faire parvenir votre curriculum vitae d'ici le 8 août, 2003 à l'adresse suivante : EN: To apply for this position, please forward your resume no later than August 8, 2003 to: Matching value: 37%	No corresponding fuzzy matching results No corresponding fuzzy matching results
15) Traducteur - traductrice (français).	FR: Traducteur-Réviseur-Terminologue EN: Translator-Editor-Terminologist Matching value: 56%	FR: Traducteur ou traductrice conseil, EN: Translator/Language Advisor, Matching value: 67%
16) Date limite :	FR: Date limite : EN: Closing Date:	FR: Date limite : le 29 mars 2005 EN: Closing Date: March 29, 2005

New ST segment to be translated	Trados fuzzy match results ~ 30% matching value	Fusion fuzzy match results ~ 50% matching value
	Matching value: 100% exact match	Matching value: 68%

APPENDIX 3.2 – CIC Immigration Appeals Fuzzy Matching Results (Trados vs. Fusion Translate)

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
1) La comparution en personne n'est pas accordée sur simple demande et, aux termes du paragraphe 175(2) de la Loi sur l'immigration et la protection des réfugiés, il doit y avoir des motifs qui étayent une telle demande.	FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Match value: 35%	No fuzzy matching results for this segment No fuzzy matching results for this segment
2) Al-Gumer, Nazer Jassim c. M.C.I. (SAI TA4-11257), Néron, 16 novembre 2004.	FR: Dikken, Cathrine Dionne c. M.C.I. (SAI TA4-11579), Stein, 19 octobre 2005. EN: Dikken, Cathrine Dionne v. M.C.I. (IAD TA4-11579), Stein, October 19, 2005. Match value: 61%	FR: Dikken, Cathrine Dionne c. M.C.I. (SAI TA4-11579), Stein, 19 octobre 2005. EN: Dikken, Cathrine Dionne v. M.C.I. (IAD TA4-11579), Stein, October 19, 2005. Match value: 58%
3) La demande parrainée de résidence permanente de la partenaire conjugale	FR: L'appelant en a appelé du refus de la demande parrainée de résidence	No fuzzy matching results for this

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy-match results – 50% matching value
de l'appelant a été refusée.	permanente de sa femme. EN: The appellant appealed the refusal of the sponsored application for permanent residence of his wife. Match value: 35%	segment No fuzzy matching results for this segment
4) La Section d'appel a déterminé que l'appelant et la demandeur n'étaient pas obligés d'avoir vécu ensemble pendant un an, ou pendant une période de temps spécifique, pour correspondre à la définition de « partenaire conjugal » figurant à l'article 2 du Règlement sur l'immigration et la protection des réfugiés.	FR: Le seul doute émis par le ministre était que l'appelant et le demandeur ne correspondaient pas à la définition de partenaires conjugués pour ce qui était de la période de un an exigée à l'article 2 du Règlement sur l'immigration et la protection des réfugiés (RIPR). EN: The Minister's sole concern was that the appellant and applicant did not qualify as conjugal partners for the one-year period required by the section 2 of the IRP Regulations. Match value: 45%	No fuzzy matching results for this segment No fuzzy matching results for this segment
5) Si la cohabitation avait été exigée, il n'aurait pas été nécessaire d'inclure les unions de fait dans la Loi sur l'immigration et la protection des	FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des	No fuzzy matching results for this segment

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
réfugiés.	réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act. Matching value: 39%	No fuzzy matching results for this segment
6) Atasever, Metin c. M.C.I. (TA3-21328), Collins, 16 novembre 2004.	FR: Atienza, Eufrosino c. M.C.I. (SAI TA4-11668), Collins, 20 octobre 2005. EN: Atienza, Eufrosino v. M.C.I. (IAD TA4-11668), Collins, October 20, 2005. Matching value: 65%	FR: Atienza, Eufrosino c. M.C.I. (SAI TA4-11668), Collins, 20 octobre 2005. EN: Atienza, Eufrosino v. M.C.I. (IAD TA4-11668), Collins, October 20, 2005. Matching value: 61%
7) Après l'entrée en vigueur de la Loi sur l'immigration et la protection des réfugiés (LIPR), l'appelant n'a pas respecté certaines conditions du sursis et a été reconnu coupable de trafic de stupéfiants.	FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act. Matching value: 39%	No fuzzy matching results for this segment No fuzzy matching results for this segment

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
8) Lors de l'appel, l'appellant n'a pas contesté la validité de cette mesure.	FR: L'appellant n'a pas contesté la validité juridique du refus. EN: The appellant did not challenge the legal validity of the refusal. Matching value: 59%	No fuzzy matching results for this segment
9) Cette demande a été rejetée.	FR: La demande a été accueillie. EN: The appeal was allowed. Matching value: 67%	No fuzzy matching results for this segment FR: La demande a été accueillie. EN: The appeal was allowed. Matching value: 53%
10) Le ministre a présenté une demande d'interdiction de divulgation de renseignements aux termes de l'article 86 de la Loi sur l'immigration et la protection des réfugiés (LIPR).	FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act. Matching value: 37%	FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act. Matching value: 53%
11) L'appellant en a appelé de la décision selon laquelle il n'avait pas respecté l'obligation de résidence	FR: L'appellant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de	FR: L'appellant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
énoncée au paragraphe 28(2) de la Loi sur l'immigration et la protection des réfugiés (LIPR).	résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 64% FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 82%	résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 72%
12) Cabriole, Hector Mercado c. M.C.I. (SAI TA3-15809), Hoare, 28 septembre	FR: Naidu, Kirstena Kushil c. M.C.I. (SAI TA4-02402), Bousfield, 16 septembre	FR: Naidu, Kirstena Kushil c. M.C.I. (SAI TA4-02402), Bousfield, 16 septembre

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
2004. 13) La demande parrainée de résidence permanente de la femme de l'appelant a été refusée parce que le mariage était une relation de mauvaise foi.	2005. EN: Naidu, Kirstena Kushil v. M.C.I. (IAD TA4-02402), Bousfield, September 16, 2005. Matching value: 67% FR: Kassam, Shazmeen c. M.C.I. (SAI TA4-08926), Boire, 12 septembre 2005. EN: Kassam, Shazmeen v. M.C.I. (IAD TA4-08926), Boire, September 12, 2005. Matching value: 57% FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ». EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child."	2005. EN: Naidu, Kirstena Kushil v. M.C.I. (IAD TA4-02402), Bousfield, September 16, 2005. Matching value: 55% FR: Kassam, Shazmeen c. M.C.I. (SAI TA4-08926), Boire, 12 septembre 2005. EN: Kassam, Shazmeen v. M.C.I. (IAD TA4-08926), Boire, September 12, 2005. Matching value: 55% FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ». EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child."

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
	Matching value: 41%	Matching value: 63%
14) Il n'était pas crédible.	FR: Il n'est pas revenu depuis. EN: He had not returned since then. Matching value: 65%	No fuzzy matching results for this segment No fuzzy matching results for this segment
15) L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son conjoint de fait.	FR: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son partenaire conjugal. EN: The appellant appealed the refusal of the sponsored application for permanent residence of her conjugal partner. Matching value: 87% FR: L'appelant en a appelé du refus de la demande parrainée de résidence permanente de sa femme. EN: The appellant appealed the refusal of the sponsored application for permanent residence of his wife.	FR: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son partenaire conjugal. EN: The appellant appealed the refusal of the sponsored application for permanent residence of her conjugal partner. Matching value: 83% No fuzzy matching results for segment #2, even though Trados' result was 80%

New ST segment to be translated	Transos fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
	Matching value: 80%	
16) L'appel a été accueilli.	FR: La demande a été accueillie. EN: The appeal was allowed. Matching value: 57%	FR: L'appel a été accueilli. EN: <i>The appeal was allowed.</i> Matching value: 100% EXACT MATCH
17) L'agent des visas a estimé que l'appelante n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).	FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 64% FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des	FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 69% No fuzzy matching results for this segment

New ST segment to be translated	Trados fuzzy match results – 30% matching value	Fusion fuzzy match results – 50% matching value
	réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Matching value: 74%	No fuzzy matching results for this segment
18) Lors de l'appel, l'appelante n'a pas contesté la validité juridique de cette décision.	FR: Il n'a pas contesté la validité juridique de la décision. EN: He did not challenge the legal validity of the decision. Matching value: 59% FR: L'appelant n'a pas contesté la validité juridique du refus. EN: The appellant did not challenge the legal validity of the refusal. Matching value: 58%	No fuzzy matching results for this segment No fuzzy matching results for this segment

APPENDIX 4.1 – Text Category Fuzzy Matching Results

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
1) La comparaison en personne n'est pas accordée sur simple demande et, aux termes du paragraphe 175(2) de la Loi sur l'immigration et la protection des réfugiés, il doit y avoir des motifs qui étayent une telle demande. Matching value: 35% FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.	1) Lorsque, pour des raisons valables, un tel engagement régulier doit être fait dans des délais trop courts pour permettre une annonce préalable du poste, le recteur ou son délégué peut autoriser des exceptions à cette règle; il en avertit alors l'Association, avec raisons à l'appui, en temps opportun. Fuzzy match 1 – 83% FR: Lorsque, pour des raisons valables, un tel engagement régulier doit être fait dans des délais trop courts pour permettre une annonce préalable du poste, le recteur ou son délégué peut autoriser des exceptions à cette règle; et il en avertit l'Association. EN: If, for valid reasons, such a regular appointment must be made within a time period too short to allow for the position to be advertised in advance, the Rector or her delegate may authorize exceptions to this rule, in which case she shall inform the Association accordingly. Fuzzy match 2 – 38%

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
	FR: Lorsque, pour des raisons valables, l'engagement d'un professeur remplaçant doit se faire dans un délai trop court pour permettre l'annonce du poste aux termes de 17.1.2.1, le doyen peut autoriser des exceptions à cette règle, pourvu qu'il en informe l'Association et, à moins que les parties en décident autrement, qu'il l'annonce dans des départements analogues d'universités canadiennes. EN: If, for valid reasons, the appointment of a replacement professor must be made within a time period too short to allow for advertising pursuant to 17.1.2.1, the dean may authorize exceptions to this rule provided she informs the Association and, unless agreed to otherwise by the parties, advertises in similar Canadian university departments.
2) Al-Gumer, Nazer Jassim c. M.C.I. (SAI TA4-11257), Nérón, 16 novembre 2004. Matching value: 61% FR: Dikken, Cathrine Dionne c. M.C.I. (SAI TA4-11579), Stein,	2) Lorsque, pour des raisons valables, l'engagement d'un professeur remplaçant doit se faire dans un délai trop court pour permettre l'annonce du poste aux termes de 17.1.2.1, le doyen peut autoriser des exceptions à cette règle, pourvu qu'il en informe l'Association, qu'il annonce dans <i>Affaires universitaires</i> conformément à 17.1.2.1, et, à moins que les parties en décident autrement, qu'il annonce dans des départements analogues d'universités canadiennes. Fuzzy match 1 – 85% FR: Lorsque, pour des raisons valables, l'engagement d'un

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
19 octobre 2005. EN: Dikken, Cathrine Dionne v. M.C.I. (IAD TA4-11579), Stein, October 19, 2005.	professeur remplaçant doit se faire dans un délai trop court pour permettre l'annonce du poste aux termes de 17.1.2.1, le doyen peut autoriser des exceptions à cette règle, pourvu qu'il en informe l'Association et, à moins que les parties en décident autrement, qu'il l'annonce dans des départements analogues d'universités canadiennes. EN: If, for valid reasons, the appointment of a replacement professor must be made within a time period too short to allow for advertising pursuant to 17.1.2.1, the dean may authorize exceptions to this rule provided she informs the Association and, unless agreed to otherwise by the parties, advertises in similar Canadian university departments. Fuzzy match 2 – 34% FR: Lorsque, pour des raisons valables, un tel engagement régulier doit être fait dans des délais trop courts pour permettre une annonce préalable du poste, le recteur ou son délégué peut autoriser des exceptions à cette règle; et il en avertit l'Association. EN: If, for valid reasons, such a regular appointment must be made within a time period too short to allow for the position to be advertised in advance, the Rector or her delegate may authorize exceptions to this rule, in which case she shall inform the Association

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
3) La demande parrainée de résidence permanente de la partenaire conjugale de l'appelant a été refusée. Matching value: 35% FR: L'appelant en a appelé du refus de la demande parrainée de résidence permanente de sa femme. EN: The appellant appealed the refusal of the sponsored application for permanent residence of his wife.	accordingly. 3) Critères pour le rang I ou la promotion au rang II ou III (bibliothécaires) Fuzzy match 1 – 36% FR: Critères de promotion au rang de bibliothécaire II ou III EN: Criteria for promotion to librarian II or III
4) La Section d'appel a déterminé que l'appelant et la demandeur n'étaient pas obligés d'avoir vécu ensemble pendant un an, ou pendant une période de temps spécifique, pour correspondre à la définition de « partenaire conjugal » figurant à l'article 2 du Règlement sur l'immigration et la protection des réfugiés. Matching value: 45% FR: Le seul doute émis par le ministre était que l'appelante et le demandeur ne correspondaient pas à la définition de partenaires conjugués pour ce qui était de la période de un an exigée à l'article 2 du Règlement sur l'immigration et la protection des réfugiés (RIPR). EN: The Minister's sole concern was that the appellant and	4) Il est entendu que les primes administratives ne sont pas comprises dans le salaire de base du bibliothécaire et qu'elles sont annulées lorsque le bibliothécaire n'occupe plus une poste comportant d'importantes fonctions administratives, sauf s'il y a eu promotion à un rang supérieur ainsi que le précisent les dispositions ci-après : Fuzzy match 1 – 33% FR: Le bibliothécaire syndiqué qui est muté à un poste dont le rang est inférieur à celui du poste qu'il occupait se voit attribuer le rang inférieur, mais son salaire demeure inchangé et continuera de progresser tant qu'il ne correspondra pas au salaire maximum prévu pour le nouveau rang. EN: Where a librarian member is transferred to a position

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
applicant did not qualify as conjugal partners for the one-year period required by the section 2 of the IRP Regulations.	with rank lower than that of her former position, she shall be given the lower rank, but her salary shall not change and shall continue to progress until its reaches the maximum for the new rank.
5) Si la cohabitation avait été exigée, il n'aurait pas été nécessaire d'inclure les unions de fait dans la Loi sur l'immigration et la protection des réfugiés. Matching value: 39% FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act.	5) (d) superviser, diriger et évaluer les travaux individuels accomplis par les étudiants, par exemple sous forme de thèse ou de mémoire; Fuzzy match 1 – 91% FR: diriger et évaluer les travaux individuels accomplis par les étudiants, par exemple sous forme de thèse ou de mémoire; EN: guiding and evaluating students' individual work, such as theses and papers;
6) Atasever, Metin c. M.C.I. (TA3-21328), Collins, 16 novembre 2004. Matching value: 65% FR: Atienza, Eufrosino c. M.C.I. (SAI TA4-11668), Collins, 20 octobre 2005. EN: Atienza, Eufrosino v. M.C.I. (IAD TA4-11668), Collins,	6) Le budget annuel disponible pour financer les coûts supplémentaires occasionnés par des congés professionnels des bibliothécaires équivaldra à 80 % du maximum du poste de bibliothécaire III. Fuzzy match 1 – 64% FR: Le budget annuel disponible pour financer les coûts supplémentaires occasionnés par des congés professionnels des bibliothécaires sera, pour la durée de la convention, de 44 994 \$.

New ST segment to be translated CIC Immigration Appeals Database October 20, 2005.	New ST segment to be translated APUO Collective Agreement Database
	EN: The annual budget available for financing the additional costs arising from professional leaves for librarians shall be \$44,994 for the life of this agreement. Fuzzy match 2 – 59% FR: Le budget annuel disponible pour financer les coûts supplémentaires occasionnés par des congés professionnels des professeurs de langue sera, pour la durée de la convention, de 19 685 \$. EN: The annual budget available for financing supplementary costs arising from professional leaves of language teachers shall be \$19,685 for the life of this agreement.
7) Après l'entrée en vigueur de la Loi sur l'immigration et la protection des réfugiés (LIPR), l'appelant n'a pas respecté certaines conditions du sursis et a été reconnu coupable de trafic de stupéfiants. Matching value: 39% FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR).	

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act.	
8) Lors de l'appel, l'appelant n'a pas contesté la validité de cette mesure.	
Matching value: 59%	
FR: L'appelant n'a pas contesté la validité juridique du refus.	
EN: The appellant did not challenge the legal validity of the refusal.	
9) Cette demande a été rejetée.	
Matching value: 37%	
FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR).	
EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act.	
10) L'appelant en a appelé de la décision selon laquelle il n'avait pas respecté l'obligation de résidence énoncée au paragraphe 28(2) de la Loi sur l'immigration et la protection des réfugiés (LIPR).	
Matching value: 82%	

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. 11) Cabriole, Hector Mercado c. M.C.I. (SAI TA3-15809), Hoare, 28 septembre 2004. Matching value: 67% FR: Naidu, Kirstena Kushil c. M.C.I. (SAI TA4-02402), Bousfield, 16 septembre 2005. EN: Naidu, Kirstena Kushil v. M.C.I. (IAD TA4-02402), Bousfield, September 16, 2005. 12) La demande parrainée de résidence permanente de la femme de l'appelant a été refusée parce que le mariage était une relation de mauvaise foi. Matching value: 41% FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ».	

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child."	
13) Il n'était pas crédible. Matching value: 65% FR: Il n'est pas revenu depuis. EN: He had not returned since then.	
14) L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son conjoint de fait. Matching value: 87% FR: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son partenaire conjugal. EN: The appellant appealed the refusal of the sponsored application for permanent residence of her conjugal partner.	
15) L'appel a été accueilli. Matching value: 57% FR: La demande a été accueillie. EN: The appeal was allowed.	
16) L'agent des visas a estimé que l'appelante n'avait pas	

New ST segment to be translated CIC Immigration Appeals Database	New ST segment to be translated APUO Collective Agreement Database
respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). Matching value: 74% FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.	
17) Lors de l'appel, l'appelante n'a pas contesté la validité juridique de cette décision. Matching value: 59% FR: Il n'a pas contesté la validité juridique de la décision. EN: He did not challenge the legal validity of the decision.	

APPENDIX 5.1 – Language Factor Fuzzy Matching Results

French-to-English database	English-to-French database
1) New ST: L'appellant en a appelé de la décision selon laquelle il n'avait pas respecté l'obligation de résidence énoncée au paragraphe 28(2) de la Loi sur l'immigration et la protection des réfugiés (LIPR).	1) New ST: The appellant appealed a decision that he had failed to fulfill the residency obligation set out in section 28(2) of the Immigration and Refugee Protection Act.
Fuzzy match #1 – 82%	Fuzzy match #1 – 70%
FR: L'appellant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).	EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.
EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.	FR: L'appellant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).
Fuzzy match #2 – 64%	Fuzzy match #2 – 57%
FR: L'appellant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés.	EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.
EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under	FR: L'appellant, âgé de 77 ans, en a appelé d'une

French-to-English database	English-to-French database
section 28 of the Immigration and Refugee Protection Act. Fuzzy match #3 – 38% FR: Il n'y a aucun droit d'appel au titre du paragraphe 63(1) de la LIPR pour ce qui est du processus visant à obtenir la résidence permanente dont il est question en l'espèce. EN: There is no right of appeal under section 63(1) of IRPA with respect to this process to obtain permanent residence. Fuzzy match #4 – 35% FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act. 2) New ST: La demande parrainée de résidence permanente de la femme de l'appelant a été refusée parce que le mariage était une relation de mauvaise foi.	conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. 2) New ST: The sponsored application for permanent residence of the appellant's wife was refused on the basis that the marriage was a bad faith relationship.

French-to-English database	English-to-French database
Fuzzy match #1 – 44% FR: La demande parrainée de résidence permanente des nièces orphelines de l'appelant a été refusée parce que les demandeurs étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés. EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased.	Fuzzy match #1 – 59% EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child." FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ».
Fuzzy match #2 – 41% FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ». EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child."	Fuzzy match #2 – 58% EN: The sponsored application for permanent residence of the appellant's husband was refused pursuant to section 4 of the IRP Regulations. FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée au titre de l'article 4 du Règlement sur l'immigration et la protection des réfugiés (RIPR).
	Fuzzy match #3 – 51% EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased.

French-to-English database	English-to-French database
	FR: La demande parrainée de résidence permanente des nièces orphelines de l'appelant a été refusée parce que les demandeurs étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés. Fuzzy match #4 – 45% EN: The sponsored application for permanent residence of the appellant's widowed mother was refused because the applicant was medically inadmissible in that she suffered from osteoarthritis in both knees. FR: La demande parrainée de résidence permanente de la mère de l'appelant, qui était veuve, a été refusée parce que la demandeur était interdite de territoire pour motifs sanitaires, en ce sens qu'elle était atteinte d'arthrose dans les deux genoux.
3) New ST: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son conjoint de fait. Fuzzy match #1 – 87% FR: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son partenaire conjugal. EN: The appellant appealed the refusal of the sponsored	3) New ST: The appellant appealed the refusal of the sponsored application for permanent residence of her common-law spouse. Fuzzy match #1 – 84% EN: The appellant appealed the refusal of the sponsored application for permanent residence of her conjugal partner. FR: L'appelante en a appelé du refus de la demande

French-to-English database	English-to-French database
application for permanent residence of her conjugal partner. Fuzzy match #2 – 80% FR: L'appelant en a appelé du refus de la demande parrainée de résidence permanente de sa femme. EN: The appellant appealed the refusal of the sponsored application for permanent residence of his wife. Fuzzy match #3 – 45% FR: L'agente immigration avait rejeté la demande parrainée de résidence permanente présentée par l'époux de l'appelante. EN: The immigration officer had refused the sponsored application for permanent residence submitted by the appellant's spouse.	parrainée de résidence permanente de son partenaire conjugal. Fuzzy match #2 – 79% EN: The appellant appealed the refusal of the sponsored application for permanent residence of his wife. FR: L'appelant en a appelé du refus de la demande parrainée de résidence permanente de sa femme. Fuzzy match #3 – 60% EN: The immigration officer had refused the sponsored application for permanent residence submitted by the appellant's spouse. FR: L'agente immigration avait rejeté la demande parrainée de résidence permanente présentée par l'époux de l'appelante.
4) New ST: L'agent des visas a estimé que l'appelante n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). Fuzzy match #1 – 74% FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à	4) New ST: The visa officer found that she had failed to satisfy the residency obligation under section 28 of the Immigration and Refugee Protection Act. Fuzzy match #1 – 66% EN: The appellant appealed a determination that he had not complied with the residency obligation under

French-to-English database	English-to-French database
l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Fuzzy match #2 – 64% FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Fuzzy match #3 – 50% FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée au titre de l'article 4 du Règlement sur l'immigration et la protection des réfugiés (RIPR). EN: The sponsored application for permanent residence of the appellant's husband was refused pursuant to section 4 of the IRP Regulations.	section 28 of the Immigration and Refugee Protection Act. FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). Fuzzy match #2 – 55% EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés.

French-to-English database	English-to-French database
5) New ST: La Section d'appel a déterminé que l'appelant et la demandeur n'étaient pas obligés d'avoir vécu ensemble pendant un an, ou pendant une période de temps spécifique, pour correspondre à la définition de « partenaire conjugal » figurant à l'article 2 du Règlement sur l'immigration et la protection des réfugiés.	5) New ST: The Appeal Division found that the appellant and the applicant were not required to reside together for one year, or any specific period of time, in order to meet the definition of "conjugal partner" in section 2 of the Immigration and Refugee Protection Regulations.
Fuzzy match #1 – 45%	Fuzzy match #1 – 33%
FR: Le seul doute émis par le ministre était que l'appelante et le demandeur ne correspondaient pas à la définition de partenaires conjugués pour ce qui était de la période de un an exigée à l'article 2 du Règlement sur l'immigration et la protection des réfugiés (RIPR).	EN: The Minister's sole concern was that the appellant and applicant did not qualify as conjugal partners for the one-year period required by the section 2 of the IRP Regulations.
EN: The Minister's sole concern was that the appellant and applicant did not qualify as conjugal partners for the one-year period required by the section 2 of the IRP Regulations.	FR: Le seul doute émis par le ministre était que l'appelante et le demandeur ne correspondaient pas à la définition de partenaires conjugués pour ce qui était de la période de un an exigée à l'article 2 du Règlement sur l'immigration et la protection des réfugiés (RIPR).
Fuzzy match #2 – 35%	
FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).	
EN: The appellant appealed a determination that he had	

French-to-English database	English-to-French database
not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Fuzzy match #3 – 33% FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés. EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.	
6) New ST: Lors de l'appel, l'appelant n'a pas contesté la validité de cette mesure. Fuzzy match #1 – 59% FR: L'appelant n'a pas contesté la validité juridique du refus. EN: The appellant did not challenge the legal validity of the refusal. Fuzzy match #2 – 45% FR: Il n'a pas contesté la validité juridique de la décision.	6) New ST: He did not challenge the validity of the order on appeal. Fuzzy match #1 – 64% EN: He did not challenge the legal validity of the decision. FR: Il n'a pas contesté la validité juridique de la décision. Fuzzy match #2 – 45% EN: The appellant did not challenge the legal validity of the refusal.

French-to-English database	English-to-French database
EN: He did not challenge the legal validity of the decision.	FR: L'appelant n'a pas contesté la validité juridique du refus.
7) New ST: Le ministre a présenté une demande d'interdiction de divulgation de renseignements aux termes de l'article 86 de la Loi sur l'immigration et la protection des réfugiés (LIPR). Fuzzy match #1 – 46% FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act. Fuzzy match #2 – 40% FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée au titre de l'article 4 du Règlement sur l'immigration et la protection des réfugiés (RIPR). EN: The sponsored application for permanent residence of the appellant's husband was refused pursuant to	7) New ST: The Minister brought an application for non-disclosure of information under section 86 of the Immigration and Refugee Protection Act. NO FUZZY MATCHES RETRIEVED

French-to-English database	English-to-French database
section 4 of the IRP Regulations. Fuzzy match #3 – 37% FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR). EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act.	
8) New ST: La demande parrainée de résidence permanente présentée par les parents de l'appelant a été refusée au motif que l'appelant ne pourrait respecter son engagement d'aide et que les requérants ne pourraient subvenir eux-mêmes à leurs besoins. Fuzzy match #1 – 41% FR: La demande parrainée de résidence permanente de la mère de l'appelant, qui était veuve, a été refusée parce que la demandeur était interdite de territoire pour motifs sanitaires, en ce sens qu'elle était atteinte d'arthrose dans les deux genoux. EN: The sponsored application for permanent residence of the appellant's widowed mother was refused because the applicant was medically inadmissible in that she	8) New ST: The sponsored application for permanent residence of the appellant's parents was refused on the ground that the appellant would be unable to fulfil his undertaking of assistance and the applicants would be unable to support themselves. Fuzzy match #1 – 46% EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased. FR: La demande parrainée de résidence permanente des nièces orphelines de l'appelant a été refusée parce que les demandeurs étaient incapables de fournir des éléments de preuve selon lesquels leurs parents

French-to-English database	English-to-French database
suffered from osteoarthritis in both knees. Fuzzy match #2 – 38% FR: La demande parrainée de résidence permanente des nièces orphelines de l'appelant a été refusée parce que les demandeurs étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés. EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased. Fuzzy match #3 – 34% FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée au titre de l'article 4 du Règlement sur l'immigration et la protection des réfugiés (RIPR). EN: The sponsored application for permanent residence of the appellant's husband was refused pursuant to section 4 of the IRP Regulations. Fuzzy match #4 – 34% FR: La demande parrainée de résidence permanente du	étaient décédés. Fuzzy match #2 – 40% EN: The sponsored application for permanent residence of the appellant's husband was refused on the ground that the marriage did not comply with the Hindu Marriage Act, 1955 as the HMA does not allow marriages between Christians (like the appellant) and Sikhs (like the applicant). FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée parce que le mariage contrevenait à la Hindu Marriage Act, 1955, car cette loi ne permet pas les mariages entre chrétiens (comme l'appelante) et sikhs (comme le demandeur). Après l'audition de son appel, l'appelante a demandé à présenter des éléments de preuve selon lesquels elle était hindoue. Fuzzy match #3 – 40% EN: The sponsored application for permanent residence of the appellant's son was refused pursuant to section 1.17(9)(d) of the Immigration and Refugee Protection Regulations as the applicant was not declared as a dependent when the appellant obtained her permanent residence. FR: La demande parrainée de résidence permanente du

French-to-English database	English-to-French database
fil de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ». EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child."	fil de l'appelante a été rejetée aux termes de l'alinéa 11.7(9)d) du Règlement sur l'immigration et la protection des réfugiés (RIPR), car le demandeur n'avait pas été déclaré comme personne à charge lorsque l'appelante a obtenu la résidence permanente. Fuzzy match #4 – 36% EN: The sponsored application for permanent residence of the appellant's widowed mother was refused because the applicant was medically inadmissible in that she suffered from osteoarthritis in both knees. FR: La demande parrainée de résidence permanente de la mère de l'appelant, qui était veuve, a été refusée parce que la demandeur était interdite de territoire pour motifs sanitaires, en ce sens qu'elle était atteinte d'arthrose dans les deux genoux. Fuzzy match #5 – 36% EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child." FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ».
9) New ST: Ferri, Loreto Lorenzo c. M.C.I. (SAI T95-02891),	10) New ST: Ferri, Loreto Lorenzo v. M.C.I. (IAD T95-02891),

French-to-English database	English-to-French database
Stein, 2 novembre 2004.	Stein, November 2, 2004.
Fuzzy match #1 – 67%	Fuzzy match #1 – 71%
FR: Dikken, Cathrine Dionne c. M.C.I. (SAI TA4-11579), Stein, 19 octobre 2005.	EN: Dikken, Cathrine Dionne v. M.C.I. (IAD TA4-11579), Stein, October 19, 2005.
EN: Dikken, Cathrine Dionne v. M.C.I. (IAD TA4-11579), Stein, October 19, 2005.	FR: Dikken, Cathrine Dionne c. M.C.I. (SAI TA4-11579), Stein, 19 octobre 2005.
Fuzzy match #2 – 64%	Fuzzy match #2 – 67%
FR: Khan, Khalid c. M.C.I. (SAI TA4-08639), Stein, 1er novembre 2005.	EN: Khan, Khalid v. M.C.I. (IAD TA4-08639), Stein, November 1, 2005.
EN: Khan, Khalid v. M.C.I. (IAD TA4-08639), Stein, November 1, 2005.	FR: Khan, Khalid c. M.C.I. (SAI TA4-08639), Stein, 1er novembre 2005.
Fuzzy match #3 – 64%	Fuzzy match #3 – 65%
FR: Wahba, Amir Anis c. M.C.I. (SAI MA5-00325), Barazi, 1er novembre 2005.	EN: Naidu, Kirstena Kushil v. M.C.I. (IAD TA4-02402), Bousfield, September 16, 2005.
EN: Wahba, Amir Anis v. M.C.I. (IAD MA5-00325, Barazi, November 1, 2005.	FR: Naidu, Kirstena Kushil c. M.C.I. (SAI TA4-02402), Bousfield, 16 septembre 2005.
Fuzzy match #4 – 62%	Fuzzy match #4 – 60%
FR: Naidu, Kirstena Kushil c. M.C.I. (SAI TA4-02402), Bousfield, 16 septembre 2005.	EN: Kassam, Shazmeen v. M.C.I. (IAD TA4-08926), Boire, September 12, 2005.

French-to-English database	English-to-French database
EN: Naidu, Kirstena Kushil v. M.C.I. (IAD TA4-02402), Bousfield, September 16, 2005. Fuzzy match #5 – 56% FR: Kassam, Shazmeen c. M.C.I. (SAI TA4-08926), Boire, 12 septembre 2005. EN: Kassam, Shazmeen v. M.C.I. (IAD TA4-08926), Boire, September 12, 2005.	FR: Kassam, Shazmeen c. M.C.I. (SAI TA4-08926), Boire, 12 septembre 2005. Fuzzy match #5 – 58% EN: Atienza, Eufrosino v. M.C.I. (IAD TA4-11668), Collins, October 20, 2005. FR: Atienza, Eufrosino c. M.C.I. (SAI TA4-11668), Collins, 20 octobre 2005.

APPENDIX 6.1

EVALUATION OF FUZZY MATCHES PROPOSED BY TM SYSTEMS

Context: Translation memory (TM) tools are designed to help translators by proposing possible translation solutions based on previously translated texts. Of course, the solutions that are proposed are not always perfect. Instead, they may take the form of fuzzy matches and may require some editing. However, these fuzzy matches may at least offer some assistance to the translator.

Instructions: You will be presented with 10 different sentences in French that need to be translated into English. For each sentence, you will be given 2 or 3 possible fuzzy match solutions that have been proposed by a TM system and that may be helpful for translating the source sentence. Read the proposed solutions carefully and then RANK these proposals from “most helpful” to “least helpful” (where a ranking of 1 = most helpful, 2 = 2nd most helpful, etc.).

Remember: you are *not* ranking which of the proposed translations is the most accurate translation of the sentence. Please rank each proposed translation according to which fuzzy match would most *help you* translate the French sentence into English.

NOTE: In order to carry out this task, you may find it helpful to actually propose a rough translation of your own. These translations will not be evaluated in any way, but a space has been provided where you can make rough notes or jot down a draft in case this exercise is useful for helping you evaluate the usefulness of the various TM proposals.

If desired, please feel free to add any comments that you feel are relevant.

Please study the example provided before going on to complete the rest of the form.

Thank-you for agreeing to participate in this study!

Matt Fifer

EXAMPLE:

New source segment to be translated:

Vous vous intéressez à l'économie et à la finance et comptez au moins cinq années d'expérience comme traducteur dans ces domaines ou d'autres domaines pertinents?

[Space to propose your rough translation if desired]

Proposals from the translation memory system:

FR: La préférence sera accordée aux personnes qui ont au moins deux années d'expérience encadrée en traduction, ou l'équivalent en formation et en expérience.

2 [your ranking]

EN: Preference will be given to applicants with at least two years of supervised experience in translation, or equivalent in training and experience.

FR: Vous avez au moins deux années d'expérience comme traducteur dans un ou plusieurs des domaines suivants :

1 [your ranking]

EN: As the ideal candidate, you are a translator with a minimum of two years experience in at least one of the following fields:

Additional comments:

SENTENCE #1:

L'appelant en a appelé de la décision selon laquelle il n'avait pas respecté l'obligation de résidence énoncée au paragraphe 28(2) de la Loi sur l'immigration et la protection des réfugiés (LIPR).

--

Proposals from the translation memory system:

FR: Il n'y a aucun droit d'appel au titre du paragraphe 63(1) de la LIPR pour ce qui est du processus visant à obtenir la résidence permanente dont il est question en l'espèce.

3 (38%)

EN: There is no right of appeal under section 63(1) of IRPA with respect to this process to obtain permanent residence.

FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).

1 (82%)

EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.

FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés.

2 (64%)

EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.

Additional comments:

SENTENCE #2:

La demande parrainée de résidence permanente de la femme de l'appelant a été refusée parce que le mariage était une relation de mauvaise foi.

Proposals from the translation memory system:

FR: La demande parrainée de résidence permanente du fils de l'appelant a été rejetée parce que ce dernier ne correspondait pas à la définition d'« enfant à charge ».

1 (44%)

EN: The sponsored application for permanent residence of the appellant's son was refused because he did not meet the definition of "dependent child."

FR: La demande parrainée de résidence permanente des nièces orphelines de l'appelant a été refusée parce que les demandeuses étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés.

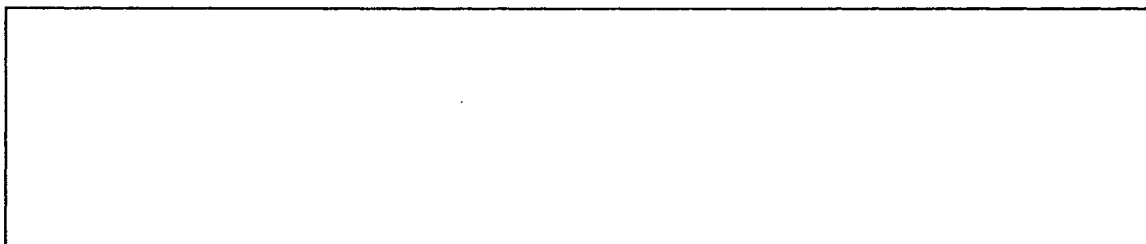
2 (41%)

EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased.

Additional comments:

SENTENCE #3:

L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son conjoint de fait.



Proposals from the translation memory system:

2 (80%)

FR: L'appelant en a appelé du refus de la demande parrainée de résidence permanente de sa femme.

EN: The appellant appealed the refusal of the sponsored application for permanent residence of his wife.

FR: L'agente immigration avait rejeté la demande parrainée de résidence permanente présentée par l'époux de l'appelante.

3 (45%)

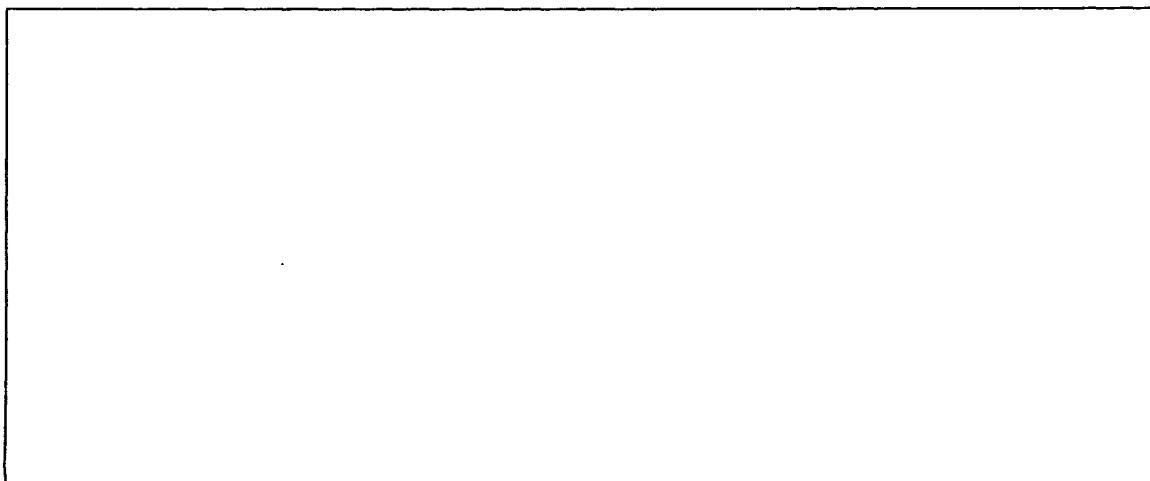
EN: The immigration officer had refused the sponsored application for permanent residence submitted by the appellant's spouse.

1 (87%)

FR: L'appelante en a appelé du refus de la demande parrainée de résidence permanente de son partenaire conjugal..

EN: The appellant appealed the refusal of the sponsored application for permanent residence of her conjugal partner.

Additional comments:



SENTENCE #4:

L'agent des visas a estimé que l'appelante n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).

Proposals from the translation memory system:

FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).

1 (74%)

EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.

FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée au titre de l'article 4 du Règlement sur l'immigration et la protection des réfugiés (RIPR).

3 (50%)

EN: The sponsored application for permanent residence of the appellant's husband was refused pursuant to section 4 of the IRP Regulations.

FR: L'appelant, âgé de 77 ans, en a appelé d'une conclusion selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés.

2 (64%)

EN: The 77-year-old appellant appealed a finding that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.

Additional comments:

SENTENCE #5:

Lors de l'appel, l'appelant n'a pas contesté la validité de cette mesure.

--

Proposals from the translation memory system:

2 (45%)

FR: Il n'a pas contesté la validité juridique de la décision.

EN: He did not challenge the legal validity of the decision.

FR: L'appelant n'a pas contesté la validité juridique du refus.

1 (56%)

EN: The appellant did not challenge the legal validity of the refusal.

Additional comments:

--

SENTENCE #6:

Le ministre a présenté une demande d'interdiction de divulgation de renseignements aux termes de l'article 86 de la Loi sur l'immigration et la protection des réfugiés (LIPR).

Proposals from the translation memory system:

FR: L'appelant en a appelé d'une décision selon laquelle il n'avait pas respecté l'obligation de résidence prévue à l'article 28 de la Loi sur l'immigration et la protection des réfugiés (LIPR).

1 (46%)

EN: The appellant appealed a determination that he had not complied with the residency obligation under section 28 of the Immigration and Refugee Protection Act.

FR: Le temps qu'il a passé à ce Centre constituait une interruption de ses études, qui n'était pas permise par la Loi sur l'immigration et la protection des réfugiés (LIPR).

3 (37%)

EN: His time at the Centre constituted a gap in his studies which was not permitted by the Immigration and Refugee Protection Act.

FR: La demande parrainée de résidence permanente du mari de l'appelante a été refusée au titre de l'article 4 du Règlement sur l'immigration et la protection des réfugiés (RIPR).

2 (40%)

EN: The sponsored application for permanent residence of the appellant's husband was refused pursuant to section 4 of the IRP Regulations.

Additional comments:

SENTENCE #7:

La demande parrainée de résidence permanente présentée par les parents de l'appelant a été refusée au motif que l'appelant ne pourrait respecter son engagement d'aide et que les requérants ne pourraient subvenir eux-mêmes à leurs besoins.

Proposals from the translation memory system:

FR: La demande parrainée de résidence permanente des nièces orphelines de l'appelant a été refusée parce que les demandeuses étaient incapables de fournir des éléments de preuve selon lesquels leurs parents étaient décédés.

2 (38%)

EN: The sponsored application for permanent residence of the appellant's orphaned nieces was refused on the basis that the applicants were unable to provide evidence that their parents were deceased.

FR: La demande parrainée de résidence permanente de la mère de l'appelant, qui était veuve, a été refusée parce que la demandeuse était interdite de territoire pour motifs sanitaires, en ce sens qu'elle était atteinte d'arthrose dans les deux genoux.

1 (41%)

EN: The sponsored application for permanent residence of the appellant's widowed mother was refused because the applicant was medically inadmissible in that she suffered from osteoarthritis in both knees.

Additional comments:

SENTENCE #8:

Maîtrise du français et de l'anglais, parlé et écrit.

--

Proposals from the translation memory system:

2 (56%)

FR: Excellente connaissance de l'anglais et du français.

EN: Excellent command of English and French.

FR: La connaissance du français et de l'anglais est essentielle.

1 (60%)

EN: Knowledge of English and French is essential.

3 (38%)

FR: Traducteur ou traductrice conseil,
de l'anglais au français

EN: Translator/Language Advisor,
English to French

Additional comments:

--

SENTENCE #9:

Au moins 2 années d'expérience comme réviseur en plus de 5 années en traduction supervisée

Proposals from the translation memory system:

2 (34%) FR: Vous avez au moins deux années d'expérience comme traducteur dans un ou plusieurs des domaines suivantes :

EN: As the ideal candidate, you are a translator with a minimum of two years experience in at least one of the following fields:

FR: Au moins 1 à 3 ans d'expérience en traduction en milieu commercial, de préférence en assurances, services financiers

1 (36%)

EN: Minimum of 1-3 years experience as a translator within a business setting, preferably in insurance, financial services

Additional comments:

SENTENCE #10:

Expérience en rédaction, révision et en traduction

--

Proposals from the translation memory system:

- | | |
|----------------|--------------------------------------|
| <u>1 (37%)</u> | FR: Expérience en révision |
| | EN: Previous revision experience |
| | FR: Vaste expérience en traduction |
| <u>2 (34%)</u> | EN: Extensive translation experience |

Additional comments:

--

APPENDIX 6.2

Chapter 6 – The Fuzzy Factor – Participant Comments

Segment 1

Participant 3: Name of the Act and wording for “selon laquelle” helpful.

Participant 4: The last two are almost equally helpful in terms of terminology and sentence structure, but the third requires slightly more editing. The main advantage of the first is that it confirms the IRPA acronym, but this is nothing that can’t be accurately guessed from the full name of the Act, which is provided in the other two. Also, that small advantage is more than offset by the misleading translation of “paragraphe” as “section,” when it should be “subsection.”

Participant 5: #1 because long clause almost exactly the same to the example sentence. #2 because has the term “paragraphe” which, as a legal term, likely has only one correct translation. #3 only bit this sentence provides what present in the other two is “résidence” which is an easy translation.

Participant 6: Second example a near 100% match – easily the most helpful.

Segment 2

Participant 1: I find these two matches to be pretty close as far as usefulness goes. The only reason I chose the second pair as most useful because of one word: “refusé,” which appears in the text to be translated, vs. “rejetée.”

Participant 4: Both are equally helpful in terms of terminology and sentence structure. The first is only slightly more helpful because it is shorter and requires less editing.

Participant 5: Look at the same bits for the two translations (first clause) but the second sounds better in English (“on the basis that” vs. “because”)

Participant 6: While both texts are helpful in a similar fashion (i.e. the first part of the fuzzy match corresponds to the first part of the source segment), the second text contains a lot of additional information that is not helpful in this context, meaning the information would need to be filtered out, adding extra work to the translation process.

Segment 3

Participant 3: Although I ranked these 1, 2, 3, all 3 were almost equally helpful.

Participant 1: The difference between 1 and 2 to me is that the English segment already contains the pronoun “her” so there would be less typing, although I’m not sure how much of a difference that would make in the big scheme of things.

Participant 4: All these are helpful for terminology, but the first and last also provide the sentence structure. The last takes slightly less time to edit

because of the feminine, although further research would be needed to determine how “conjoint de fait” is actually translated in the IRPA. I doubt that it would be conjugal partner, even if the meaning is just about the same.
Participant 5: #1 & #2 are exactly the same in terms of usefulness, but #2 has a punctuation error (two periods at the end of the sentence)
Participant 6: 1 and 2 are nearly identical, making their helpfulness practically identical. However, the first match was selected because it is shorter.

Segment 4

Participant 2: None of these are particularly helpful!
Participant 4: The first and third sentences are almost equally helpful. Neither provides the acronym IRPA (although it is easy to guess), despite the fact that LIPR is included in the first French fuzzy match. The last sentence is only slightly less helpful because of the extra info (77-year-old) that needs deleting.
Participant 5: #1 and #2 are similar but #2 has a longer different first clause than #1 making it seem potentially not as good a match upon a first reading.

Segment 5

Participant 2: I gave these sentences a ranking because that's what the instructions say to do, but I find both of these sentences equally helpful. I ranked the first one as the best for no reason other than because it is a bit shorter and faster to read.
Participant 4: Both sentences give me “challenge,” which is the information I need. The second is slightly more helpful because the full noun is provided instead of the pronoun.
Participant 6: As “the appellant” is already indicated in the second matched text, it would be more helpful as there would ultimately be fewer modifications to the target.

Segment 6

Participant 3: None of these matches were particularly helpful, especially after the previous sentences (I know the name of the Act after doing them).
Participant 1: If I was a translator in this subject field, I probably wouldn't use any of these matches and instead simply translate the sentence from scratch.
Participant 4: Close tie between the last two for second place. The middle sentence has the advantage of giving the name of the act, but the last sentence provides at least a hint of it, as well as “application” and “pursuant to,” both of which could be helpful.
Participant 5: #1 because lists same law. #2 ditto.

Segment 7

Participant 2: Again, I did not find either one much more helpful than the other.

Participant 4: Actually, I am tempted to give these identical rankings. Each gives exactly the same useful terms, and “on the basis that” and “because” are both quite appropriate in this context. The only minuscule advantage in the first is that it already has the word “unable” in it.

Participant 5: “On the basis that” over “because.”

Participant 6: As discussed in an earlier example, since the second match contains a large amount of additional information preference was given to the first for its brevity.

Segment 8

Participant 1: I’m afraid none of these three matches would be very useful to me.

Participant 4: Neither of the last two is useful at all (assuming that no translator needs help translating the words “anglais” and “français”).

“Excellent command,” on the other hand, could be quite useful.

Segment 9

Participant 1: Are translation memoParticipant 8s capable of recognizing “2” and “deux” as a match?

Participant 4: The second sentence is structured more similarly to the new sentence, and it inspires/confirms the modulation from “en traduction” to “as a translator,” which is more helpful despite the need to change the numbers.

Participant 5: #1 – same telegraphic style

Segment 10

Participant 2: Neither one of these is very helpful, it would be faster to write my own translation than to use and modify either of these matches.

Participant 4: The first sentence has the advantage of making one pause to consider whether it would be appropriate to add “previous,” a possibility not prompted by the new source sentence itself.

Participant 5: #2 is tricky because I would translate “revision” with “editing.” Suggests close translations elsewhere in the text that match was pulled from.

Participant 8: It was somewhat difficult to choose between the two options, but I ranked the 2nd option as being the most helpful because I felt that “expeParticipant 8nce” already included the concept of “previous” so it could be redundant.