

Lynne Bowke

# Computer-Aided Translation Technology

A Practical Introduction

University of Ottawa Press

LYNNE BOWKER

# Computer-Aided Translation Technology: A Practical Introduction

Didactics of Translation Series  
University of Ottawa Press

© University of Ottawa Press, 2002

ISBN 0-7766-3016-4 (cloth)

ISBN 0-7766-0538-0 (paper)

Printed in Canada



Printed on acid-free paper

---

### National Library of Canada Cataloguing in Publication Data

Bowker, Lynne, 1969–

Computer-aided translation technology: a practical  
introduction

(Didactics of translation series)

Includes bibliographical references and index.

ISBN 0-7766-3016-4 (bound) – ISBN 0-7766-0538-0 (pbk.)

1. Machine translating. 2. Translating machines. I. Title.

II. Series

P308.B69 2002 418'.02'0285 C2001-904318-X

---

University of Ottawa Press gratefully acknowledges the support extended to its publishing program by the Canada Council, the Department of Canadian Heritage, and the University of Ottawa.

“All rights reserved. No part of this publication may be reproduced or transmitted in any form or any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.”

This book has been published with the help of a grant from the University of Ottawa Faculty of Arts.

## DIDACTICS OF TRANSLATION SERIES

Catering to the needs of students in schools of translation and interpretation, the textbooks published in this series are also very helpful to professional translators and interpreters who wish to improve their technique. The series' titles cover various fields in the discipline such as general translation and specialized translation as well as editing, writing, and lexicology for translators. Works that analyse the discipline from a more theoretical or practical point of view can be found in the "Perspectives on Translation" series. Both series welcome manuscripts written in either English or French.

### Advisory committee

Jean Delisle, Series Director, University of Ottawa  
Marie-Christine Aubin, Collège universitaire de Saint-Boniface  
Michel Ballard, Université d'Artois, Arras  
Annie Brisset, University of Ottawa  
Monique C. Cormier, Université de Montréal  
Hannelore Lee-Jahnke, Université de Genève  
Daniel Siméoni, York University  
Lawrence Venuti, Temple University, Philadelphia  
Luise von Flotow, University of Ottawa  
Agnès Whitfield, York University

### In the same series

Jean Delisle, *La traduction raisonnée: Manuel d'initiation à la traduction professionnelle de l'anglais vers le français*, 1993  
Jean Delisle, *La traduction raisonnée: Livre du maître*, 1993  
Jean Delisle et Judith Woodsworth (dir.), *Les Traducteurs dans l'histoire*, 1995  
Allison Beeby Lonsdale, *Teaching Translation from Spanish to English. Worlds beyond Words*, 1996

# Acknowledgments

I am grateful to many people for their support and encouragement during the production of this volume. In particular, my thanks are owed to Jean Delisle and Ingrid Meyer of the University of Ottawa, and to Dorothy Kenny, Jennifer Pearson, and Andrew Way of Dublin City University, who offered valuable feedback on earlier versions of this work. Any remaining errors or omissions are, of course, entirely my own.

I would also like to express my appreciation to my former translation technology students, in both Dublin and Ottawa, for their stimulating questions and discussions on many aspects of computer-aided translation technology.

Thanks are also due to the Faculty of Arts of the University of Ottawa for its support of this publication.

Finally, this project could not have been achieved without the support of my family – Keith, Joyce, Lisa, and Peter. This book is dedicated to them.

# Abbreviations

This is a list of abbreviations used throughout this book. Appendix A contains a glossary that explains many key terms relating to translation technology.

ASCII	American Standard Code for Information Interchange
CAT	Computer-aided translation
CD-ROM	Compact disk read only memory
DBCS	Double-byte character set
EBMT	Example-based machine translation
HAMT	Human-assisted machine translation
HTML	HyperText Markup Language
ISO	International Organization for Standardization
KWIC	Key word in context
LISA	Localization Industry Standards Association
MAHT	Machine-assisted human translation
MARTIF	Machine Readable Terminology Interchange Format
MAT	Machine-assisted translation
MB	Megabyte(s)
MI	Mutual information
MIME	Multipurpose Internet Mail Extensions
MT	Machine translation
OCR	Optical character recognition
OSCAR	Open Standards for Container/Content Allowing Reuse
PDF	Portable document format
RAM	Random-access memory
RTF	Rich text format
TBX	Term Base eXchange

## xx Abbreviations

TM	Translation memory
TMS	Terminology-management system
TMX	Translation Memory eXchange
WWW	World Wide Web
XML	eXtensible Markup Language

# Contents

LIST OF TABLES	xi
LIST OF FIGURES	xiii
ACKNOWLEDGMENTS	xvii
ABBREVIATIONS	xix

## **0. Introduction** 3

- 0.1 Aim 4
- 0.2 Audience 5
- 0.3 Contents and coverage 5
- 0.4 Outline 9

## **1. Why Do Translators Need to Learn about Technology?** 11

- 1.1 Translation technology in the classroom:  
Laying the foundation for new types of  
investigations 15
  - 1.1.1 Exploring the impact of technology on translation  
pedagogy 15
  - 1.1.2 Investigating human-machine interaction 16
  - 1.1.3 Learning to evaluate technology 17
  - 1.1.4 Examining how tools can change conventional  
practices 17
  - 1.1.5 Producing data for empirical investigations 18
  - 1.1.6 Reinforcing basic translation skills 20
- Key points 20
- Further reading 21



<b>2. Capturing Data in Electronic Form</b>	<b>22</b>
2.1 Scanning and optical character recognition	23
2.1.1 Scanning	23
2.1.2 Optical character recognition	26
2.1.2.1 Factors affecting the accuracy of OCR	26
2.1.3 Benefits and drawbacks of scanning and OCR	28
2.1.3.1 Injury	28
2.1.3.2 Time	28
2.1.3.3 Quality	28
2.1.3.4 Languages and file formats	29
2.1.3.5 Economic aspects	29
2.2 Voice recognition	30
2.2.1 Types of voice-recognition technology	31
2.2.2 Tips for improving the accuracy of voice-recognition technology	33
2.2.3 Benefits and drawbacks of voice-recognition technology	34
2.2.3.1 Injury	34
2.2.3.2 Time	35
2.2.3.3 Quality	35
2.2.3.4 Languages and file formats	36
2.2.3.5 Integration with other tools	36
2.2.3.6 Economic aspects	36
2.3 File formats and file conversion	37
Key points	39
Further reading	42
<b>3. Corpora and Corpus-Analysis Tools</b>	<b>43</b>
3.1 Electronic corpora	44
3.1.1 Some different types of electronic corpora	45
3.2 Corpus-analysis tools	46
3.2.1 Word-frequency lists	47
3.2.1.1 Lemmatized lists	49
3.2.1.2 Stop lists	52
3.2.2 Concordancers	53
3.2.2.1 Monolingual concordancers	53
3.2.2.2 Bilingual concordancers	55
3.2.3 Collocations	64
3.3 Annotation and mark-up	68

3.4	Benefits and drawbacks of working with corpus-analysis tools	70
3.4.1	Frequency data	70
3.4.2	Context	71
3.4.3	Availability and copyright	71
3.4.4	Pre-processing	72
3.4.5	Speed and information-retrieval issues	72
3.4.6	Character sets and language-related difficulties	74
3.4.7	Economic aspects	75
	Key points	75
	Further reading	76
<b>4.</b>	<b>Terminology-Management Systems</b>	<b>77</b>
4.1	Storage	78
4.2	Retrieval	79
4.3	Active terminology recognition and pre-translation	81
4.4	Term extraction	82
4.4.1	Linguistic approach	83
4.4.2	Statistical approach	84
4.5	Additional features	86
4.6	Benefits and drawbacks of working with a TMS	86
4.6.1	Speed and flexibility	86
4.6.2	Quality	87
4.6.3	Changing the nature of the task	87
4.6.4	Shareability of information: networking, file formats, and standards	88
4.6.5	Character sets and language-related difficulties	89
4.6.6	Economic aspects	90
	Key points	90
	Further reading	91
<b>5.</b>	<b>Translation-Memory Systems</b>	<b>92</b>
5.1	How does a TM system work?	94
5.1.1	Segmentation	94
5.1.2	Matches	95
5.1.2.1	Exact matches	96
5.1.2.2	Full matches	98
5.1.2.3	Fuzzy matches	98
5.1.2.4	Term matches	101

5.1.2.5	Sub-segment matches	103
5.1.2.6	No matches	106
5.1.2.7	Limitations of existing matching algorithms	106
5.2	Creating a TM	107
5.2.1	Interactive translation	108
5.2.2	Post-translation alignment	109
5.3	Working with an existing TM	111
5.3.1	Interactive mode	111
5.3.2	Batch mode	112
5.4	Texts that are suitable for use with a TM	112
5.4.1	Texts containing internal repetitions	112
5.4.2	Revisions	113
5.4.3	Recycled texts	113
5.4.4	Updates	114
5.5	Benefits and drawbacks of working with a TM	114
5.5.1	Time	115
5.5.2	Quality	116
5.5.3	Electronic form	118
5.5.4	File formats, filters, and standards	118
5.5.5	Character sets and language-related difficulties	119
5.5.6	Attitudes	120
5.5.7	Rates of pay	121
5.5.8	Ownership	122
5.5.9	Integration with other tools	123
5.5.9.1	Terminology-management systems	123
5.5.9.2	Bilingual concordancers	124
5.5.9.3	Machine-translation systems	124
5.5.10	Economic aspects	125
	Key points	127
	Further reading	127
<b>6.</b>	<b>Other New Technologies and Emerging Trends</b>	<b>129</b>
6.1	New attitudes toward translation and translators	129
6.2	New types of translation work generated by technology	131
6.3	New technology generated by new types of translation work	132
6.4	Conditions required to ensure the continued success of CAT tools	136
6.5	Future developments	137

Key points 139

Further reading 140

APPENDIX A: GLOSSARY 141

APPENDIX B: SOME COMMERCIALY AVAILABLE CAT TOOLS 157

B.1 OCR software 157

B.2 Voice-recognition software 159

B.3 Conversion software 162

B.4 Corpus-analysis software 164

B.5 Terminology-management and translation-memory  
systems 166

B.6 Localization and Web-page translation tools 170

B.7 Word-counting tools 172

B.8 Cost/benefit estimators 173

REFERENCES 175

INDEX 183

# List of tables

- 0.1 An overview of some different types of technology used in translation
- 2.1 Different file formats and some associated conversion problems
- 3.1 Collocates appearing up to three positions to the left and right of the node "infected"
- 3.2 Sample patterns in which the node "infected" is juxtaposed with collocates or is separated from collocates by one or two intervening words
- 3.3 Two examples of part-of-speech tagging for the sentence "Scan for viruses regularly"
- 4.1 Sample term records retrieved using fuzzy matching
- 4.2 Sample hit lists retrieved for different search patterns
- 4.3 Automatic replacement of source-text terms with translation equivalents found in a term base
- 5.1 Some different types of segmentation
- 5.2 Examples of segments that will not be retrieved as exact matches
- 5.3 Example of an exact match retrieved from a TM
- 5.4 Examples from a TM that uses character matching
- 5.5 Term matches retrieved by a TM system working in conjunction with a term base
- 5.6 Examples of shortcomings in TM systems
- 5.7 Sample translation units in which the French translations are longer than the English source segments
- 6.1 Contents of a source file, translatable text extracted from that file, and a reference file containing place-holders

# List of tables

- 0.1 An overview of some different types of technology used in translation
- 2.1 Different file formats and some associated conversion problems
- 3.1 Collocates appearing up to three positions to the left and right of the node "infected"
- 3.2 Sample patterns in which the node "infected" is juxtaposed with collocates or is separated from collocates by one or two intervening words
- 3.3 Two examples of part-of-speech tagging for the sentence "Scan for viruses regularly"
- 4.1 Sample term records retrieved using fuzzy matching
- 4.2 Sample hit lists retrieved for different search patterns
- 4.3 Automatic replacement of source-text terms with translation equivalents found in a term base
- 5.1 Some different types of segmentation
- 5.2 Examples of segments that will not be retrieved as exact matches
- 5.3 Example of an exact match retrieved from a TM
- 5.4 Examples from a TM that uses character matching
- 5.5 Term matches retrieved by a TM system working in conjunction with a term base
- 5.6 Examples of shortcomings in TM systems
- 5.7 Sample translation units in which the French translations are longer than the English source segments
- 6.1 Contents of a source file, translatable text extracted from that file, and a reference file containing place-holders

# List of figures

- 2.1 A text divided into large pixels
- 2.2 The letter "H" divided into pixels
- 2.3 Sample texts of differing quality
- 3.1 A word-frequency list showing types on the left and tokens on the right
- 3.2 Word-frequency lists sorted in order of appearance in the corpus, in descending order and ascending order
- 3.3 Word-frequency lists sorted in alphabetical order, in descending order and ascending order
- 3.4 Word-frequency lists sorted in order of frequency, in descending order and ascending order
- 3.5 Twenty most frequently occurring types in an unlemmatized corpus
- 3.6 A lemmatized word list
- 3.7 Twenty most frequently occurring types in a corpus for which a stop list has been implemented
- 3.8 A sample extract from an index indicating the words contained in the corpus and the location of each occurrence
- 3.9 A KWIC display of the concordances retrieved for the search pattern "virus"
- 3.10 A KWIC display of the concordances retrieved for the search pattern "virus" sorted in alphabetical order according to the word immediately preceding "virus"
- 3.11 A KWIC display of the concordances retrieved for the search pattern "virus" sorted in alphabetical order according to the word immediately following "virus"

- 3.12 A KWIC display of the concordances retrieved using the wildcard search pattern "virus\*"
- 3.13 A KWIC display of the concordances retrieved using a context search in which "drive" must occur within a five-word span of "disk"
- 3.14 Display showing paragraphs aligned side by side, with the search pattern "virus" entered using French as the search language
- 3.15 Display showing sentences aligned one above the other, with the wildcard search pattern "clean\*" entered using English as the search language
- 3.16 A bilingual KWIC display in which source- and target-language concordances are displayed in separate windows
- 3.17 A bilingual KWIC display in which source- and target-language concordances have been sorted independently
- 3.18 Concordances retrieved using the bilingual query "drive" / "lecteur"
- 3.19 Concordances that would not be retrieved using the bilingual query "drive" / "lecteur"
- 4.1 TMS term record with a fixed set of predefined fields
- 4.2 TMS term record with free entry structure
- 4.3 A short text that has been processed using a linguistic approach to term extraction
- 4.4 A slightly modified version of the text that has been processed using a linguistic approach to term extraction
- 4.5 A short text that has been processed by a statistical term-extraction tool using a minimum frequency threshold of two
- 4.6 A slightly modified version of the text that has been processed by a statistical term-extraction tool using a minimum frequency threshold of two
- 4.7 A sample term record in which multiple forms of the term have been recorded
- 5.1 Display of translation units
- 5.2 A new source text segment and the matching translation unit stored in the TM
- 5.3 A fuzzy match retrieved from a TM
- 5.4 A TM segment retrieved using a high sensitivity threshold
- 5.5 Example of a TM segment retrieved using a low sensitivity threshold
- 5.6 Multiple fuzzy matches retrieved and ranked



- 5.7 A fuzzy match based on overall similarity
- 5.8 A sub-segment match
- 5.9 Different sub-segment matches can be combined to provide a translator with helpful suggestions
- 5.10 A bilingual concordance retrieved from a TM for the search pattern “not valid”
- 6.1 A sample of source code for a text that has been encoded in HTML
- 6.2 The text as it would appear on the World Wide Web when viewed using a browser
- 6.3 Examples of meta tags containing text strings that need to be translated