

-ing Words in RBMT: Multilingual Evaluation and Exploration of Pre- and Post-processing Solutions

Nora Aranberri Monasterio, M.Sc.

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

School of Applied Language and Intercultural Studies

Supervisor: Dr. Sharon O'Brien

November 2009

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: _____ (Candidate) ID No.: _____ Date: _____

ACKNOWLEDGEMENTS

The Enterprise Ireland Innovation Partnerships Programme Scholarship in conjunction with Symantec Ltd. afforded me the opportunity to perform research in both an academic and an industrial setting. Within these contexts, I would like to single out some people to whom I am particularly indebted.

First of all I would like to thank my academic supervisor Dr. Sharon O'Brien and my industrial supervisor Dr. Fred Hollowood for their unconditional support, expert guidance, constant presence and endless patience. Especially to Sharon, thanks for questioning all steps from an academic point of view. To Fred, thanks for the whiteboard discussions to help me make sense of the torrents of data.

I am also grateful to Prof. Jenny Williams for being an official supervisor for the first year, with a very much appreciated comeback for the last sprint. Also, I would like to thank Dr. Johann Roturier for his day-to-day support, expert advice and infinite patience.

I would also like to thank the whole Localisation Department at Symantec, and in particular the Global Language Services group (nobody managed to escape the -ings, did you?!), for the support provided by granting me access to resources and making staff available when required, and also for making me feel part of the group.

I would also like to acknowledge the guidance provided by Dr. Sabine Lehmann from acrolinx during the controlled language rule writing stage, and by Dr. Lynn Killen from the School of Computer Applications on statistics throughout the dissertation. A very special thanks goes to Julia Schinharl and Midori Tatsumi for assisting in the analysis of German and Japanese.

To all the friends from the postgrad's a big thanks for the support and for providing the so-important distraction!

Azkenik, aita eta amari, Skypeko errege eta erreginari, egunean hogeita lau ordu animatzeko prest egoteagatik musu handi-handi bat.

TABLE OF CONTENTS

ABSTRACT	xvi
INTRODUCTION.....	1
RESEARCH OBJECTIVES	1
<i>Background Research.....</i>	<i>1</i>
<i>Research Questions</i>	<i>3</i>
INDUSTRY-ACADEMIA RESEARCH COLLABORATION.....	3
<i>Symantec.....</i>	<i>4</i>
<i>Symantec Localisation Workflow</i>	<i>5</i>
<i>Relevance of Symantec-based Research.....</i>	<i>6</i>
STRUCTURE OF THE DISSERTATION.....	7
CHAPTER 1: LITERATURE REVIEW	10
1.1 DEFINING -ING WORDS	10
1.2 MACHINE TRANSLATION SYSTEMS	12
1.2.1 <i>Rule-based Machine Translation.....</i>	<i>14</i>
1.3 -ING WORDS AND (MACHINE)TRANSLATABILITY	18
1.4 MACHINE TRANSLATION EVALUATION.....	22
1.4.1 <i>Human Evaluation.....</i>	<i>23</i>
1.4.2 <i>Automatic Evaluation</i>	<i>26</i>
1.5 MACHINE TRANSLATION IMPROVEMENT	29
1.5.1 <i>Pre-processing.....</i>	<i>29</i>
1.5.2 <i>Post-processing</i>	<i>43</i>
1.6 CHAPTER SUMMARY	49
CHAPTER 2: EVALUATION METHODOLOGY	51
2.1 BUILDING A CORPUS	51
2.1.1 <i>Corpus-based Approach.....</i>	<i>51</i>
2.1.2 <i>Pilot Study</i>	<i>56</i>
2.1.3 <i>Classification.....</i>	<i>59</i>
2.2 HUMAN EVALUATION.....	72
2.2.1 <i>Evaluation</i>	<i>72</i>
2.2.2 <i>Analysing French and Spanish.....</i>	<i>85</i>
2.2.3 <i>Analysing Japanese and German</i>	<i>87</i>
2.3 AUTOMATIC EVALUATION	88
2.3.1 <i>Automatic Metrics</i>	<i>88</i>
2.3.2 <i>Considerations for the Experiment.....</i>	<i>96</i>
2.3.3 <i>Experiment Set-up</i>	<i>97</i>
2.4 CHAPTER SUMMARY	101
CHAPTER 3: DATA ANALYSIS I.....	103
3.1 HUMAN EVALUATION	103
3.1.1 <i>French and Spanish.....</i>	<i>105</i>
3.1.2 <i>Japanese and German</i>	<i>149</i>
3.1.3 <i>Summary for the Human Evaluation Analysis.....</i>	<i>171</i>
3.2 AUTOMATIC METRICS	172
3.2.1 <i>Summary for the Automatic Evaluation Analysis</i>	<i>179</i>
3.3 CHAPTER SUMMARY	179

CHAPTER 4: EXPLORATION OF APPROACHES FOR IMPROVING THE MACHINE TRANSLATION OUTPUT OF -ING WORDS.....	181
4.1 PRE-PROCESSING.....	181
4.1.1 <i>Controlled Language</i>	181
4.1.2 <i>Automatic Source Re-writing</i>	197
4.2 POST-PROCESSING.....	210
4.2.1 <i>Global Search & Replace</i>	210
4.2.2 <i>Statistical Post-editing</i>	215
4.3 CHAPTER SUMMARY	219
CHAPTER 5: DATA ANALYSIS II	222
5.1 CONTROLLED LANGUAGE	222
5.1.1 <i>Evaluation Set-up</i>	223
5.1.2 <i>Evaluation Results</i>	225
5.1.3 <i>Deployment into the Workflow</i>	227
5.2 AUTOMATIC SOURCE RE-WRITING	230
5.2.1 <i>Evaluation Set-up</i>	230
5.2.2 <i>Evaluation Results</i>	231
5.2.3 <i>Deployment into the Workflow</i>	234
5.3 GLOBAL SEARCH & REPLACE	234
5.3.1 <i>Evaluation Set-up</i>	235
5.3.2 <i>Evaluation Results</i>	235
5.3.3 <i>Deployment into the Workflow</i>	236
5.4 STATISTICAL POST-EDITING	237
5.4.1 <i>Evaluation Set-up</i>	237
5.4.2 <i>Evaluation Results</i>	237
5.4.3 <i>Deployment into the Workflow</i>	240
5.5 CHAPTER SUMMARY	241
CHAPTER 6: CONCLUSIONS	244
6.1 OBJECTIVES	244
6.2 FINDINGS.....	244
6.3 REVIEW OF METHODOLOGIES	247
6.4 FUTURE CHALLENGES AND RESEARCH OPPORTUNITIES	249
6.5 CLOSING REMARKS	250
REFERENCES.....	251
APPENDICES.....	267

APPENDICES

APPENDIX A: COMPLETE FUNCTIONAL CLASSIFICATION OF THE –ING WORDS IN THE CORPUS	268
APPENDIX B: ING EVALUATION GUIDELINES	270
APPENDIX C: QUESTIONNAIRE FOR EVALUATORS	278
APPENDIX D: LIST OF ABBREVIATIONS FOR -ING SUBCATEGORIES	280
APPENDIX E: GUIDELINES FOR THE ANALYSIS OF GERMAN AND JAPANESE -ING EVALUATION	283
APPENDIX F: CLEANING GUIDELINES FOR THE FEATURE-BASED AUTOMATIC EVALUATION	292
APPENDIX G: HUMAN EVALUATION AND AUTOMATIC METRICS CORRELATIONS	297
APPENDIX H: CL TRANSLATION EFFECT EVALUATION GUIDELINES	303
APPENDIX I: AUTOMATIC SOURCE RE-WRITING, GLOBAL S&R, & SPE TRANSLATION EFFECT EVALUATION GUIDELINES	306

LIST OF TABLES

TABLE 1.1: CLUES FOR CLASSIFYING -ING WORDS	12
TABLE 1.2: EXAMPLE OF STRUCTURAL AMBIGUITY	18
TABLE 1.3: EXAMPLE OF A PASSIVE SENTENCE WITHOUT AN EXPLICIT AGENT AND AN IMPLICIT SUBJECT SUBORDINATE CLAUSE	19
TABLE 1.4: EXAMPLE OF AN UNGRAMMATICAL USE OF IMPLICIT SUBJECTS	19
TABLE 1.5: EXAMPLES OF DIFFERENT CLASSES OF -ING WORDS	20
TABLE 1.6: TAGGING AMBIGUITY ISSUE	20
TABLE 1.7: SET OF FOUR REFERENCE TRANSLATIONS, AN MT OUTPUT AND TWO VARIATIONS WITH THE SAME BLEU SCORE FROM THE 2005 NIST MT EVALUATION (FROM CALLISON-BURCH ET AL. 2006)	28
TABLE 1.8: EXAMPLE OF NATURAL AND CONTROLLED LANGUAGE SENTENCES (TAKEN FROM QUAH 2006: 48)	40
TABLE 2.1: RESULTS FOR CORRECTNESS ACROSS LANGUAGES (PILOT PROJECT)	58
TABLE 2.2: SUMMARY OF THE MAIN FUNCTIONAL CATEGORIES OF -ING FORMS OBSERVED IN THE STUDY OF IZQUIERDO (2006)	61
TABLE 2.3: EXAMPLE OF RELEVANT AND IRRELEVANT MATCHES OF THE RULE WRITTEN TO SEARCH FOR REDUCED RELATIVE CLAUSES	64
TABLE 2.4: EXAMPLES OF THE 20% -ING WORDS NOT FOUND DURING THE SEMI-AUTOMATIC EXTRACTION	65
TABLE 2.5: EXAMPLES OF PRE-MODIFIERS	65
TABLE 2.6: EXAMPLES OF PRE-MODIFIERS WITH A PREDOMINANT ADJECTIVAL AND VERBAL FLAVOUR	66
TABLE 2.7: EXAMPLE OF A REDUCED RELATIVE CLAUSE INTRODUCED BY AN -ING WORD	66
TABLE 2.8: EXAMPLE OF NOMINAL ADJUNCT	66
TABLE 2.9: EXAMPLE OF AN ADJECTIVAL ADJUNCT	67
TABLE 2.10: BREAKDOWN OF -ING WORDS FOUND IN NOUN OR ADJECTIVE MODIFYING FUNCTION	67
TABLE 2.11: EXAMPLES OF ADVERBIAL CLAUSES WITH -ING WORDS AS HEADS	68
TABLE 2.12: EXAMPLES OF AMBIGUOUS <i>FOR</i> + -ING STRUCTURES	68
TABLE 2.13: BREAKDOWN OF -ING WORDS FOUND INTRODUCING ADVERBIAL CLAUSES DIRECTLY OR PRECEDED BY A PREPOSITION/SUBORDINATE CONJUNCTION	69
TABLE 2.14: EXAMPLES OF -ING WORDS WITH A PROGRESSIVE ASPECT FUNCTION	69
TABLE 2.15: EXAMPLES OF AMBIGUOUS INSTANCES FOR PROGRESSIVE-ASPECT -ING WORDS	70
TABLE 2.16: BREAKDOWN OF -ING WORDS FOUND INTRODUCING THE PROGRESSIVE ASPECT	70
TABLE 2.17: BREAKDOWN OF REFERENTIAL FUNCTION -ING WORDS	71
TABLE 2.18: BREAKDOWN OF -ING WORDS AT THE BEGINNING OF TITLES	72
TABLE 2.19: SAMPLE OF THE -ING WORDS OBTAINED BY APPLYING A STRATIFIED SYSTEMATIC SAMPLING METHOD.	85
TABLE 2.20: EXAMPLES SHOWING FORMATTING TRANSFER ERRORS BY THE RBMT SYSTEM	98
TABLE 2.21: EXAMPLE OF AUTOMATIC METRICS' INACCOUNTABILITY OF LONG-DISTANCE ISSUES.	100
TABLE 3.1: OVERALL HUMAN EVALUATION RESULTS FOR FRENCH	103
TABLE 3.2: OVERALL HUMAN EVALUATION RESULTS FOR GERMAN	103
TABLE 3.3: OVERALL HUMAN EVALUATION RESULTS FOR JAPANESE	104
TABLE 3.4: OVERALL HUMAN EVALUATION RESULTS FOR SPANISH	104
TABLE 3.5: CLASSIFICATION OF BEST TO WORST PERFORMING CATEGORY ACROSS LANGUAGES	104
TABLE 3.6: SUBCATEGORIES WITHIN THE CATEGORY OF TITLES	105

TABLE 3.7: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF TITLES STARTING WITH AN -ING WORD	106
TABLE 3.8: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES STARTING WITH AN -ING WORD	107
TABLE 3.9: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES WITHIN QUOTATION MARKS AT THE BEGINNING OF SENTENCE.....	107
TABLE 3.10: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES WITHIN QUOTATION MARKS EMBEDDED IN A SENTENCE	108
TABLE 3.11: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES STARTING WITH <i>ABOUT</i> FOLLOWED BY AN -ING WORD	109
TABLE 3.12: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF TITLES STARTING WITH AN -ING WORD	109
TABLE 3.13: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES STARTING WITH AN -ING WORD	110
TABLE 3.14: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES WITHIN QUOTATION MARKS AT THE BEGINNING OF SENTENCE.....	111
TABLE 3.15: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES WITHIN QUOTATION MARKS EMBEDDED IN A SENTENCE	111
TABLE 3.16: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF TITLES STARTING WITH <i>ABOUT</i> FOLLOWED BY AN -ING WORD	112
TABLE 3.17: SUBCATEGORIES WITHIN THE CATEGORY OF CHARACTERISERS	114
TABLE 3.18: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF PRE-MODIFYING -ING WORDS.....	114
TABLE 3.19: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF PRE-MODIFYING -ING WORDS	115
TABLE 3.20: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF PRE-MODIFYING -ING WORDS	116
TABLE 3.21: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF REDUCED RELATIVE CLAUSES.....	116
TABLE 3.22 ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF PRE-MODIFYING -ING WORDS	117
TABLE 3.23: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF NOMINAL ADJUNCTS.....	118
TABLE 3.24: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF NOMINAL ADJUNCTS CONTAINING -ING WORDS.....	118
TABLE 3.25: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF PRE-MODIFYING -ING WORDS.....	119
TABLE 3.26: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF PRE-MODIFYING -ING WORDS	120
TABLE 3.27: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF POST-MODIFYING -ING WORDS.....	120
TABLE 3.28 ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF POST-MODIFYING -ING WORDS.....	121
TABLE 3.29: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF REDUCED RELATIVE CLAUSES	122
TABLE 3.30: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF REDUCED RELATIVE CLAUSES.....	122
TABLE 3.31: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF NOMINAL ADJUNCTS.....	123
TABLE 3.32: ISSUES FOUND FOR THE TRANSLATION OF THE SUCATEGORY OF NOMINAL ADJUNCTS	123
TABLE 3.33: SUBCATEGORIES WITHIN THE CATEGORY OF PROGRESSIVES	125
TABLE 3.34: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF ACTIVE VOICE PRESENT CONTINUOUS TENSE.....	126

TABLE 3.35: ISSUES FOUND FOR THE TRANSLATION OF THE SUBCATEGORY OF ACTIVE VOICE	
PRESENT CONTINUOUS TENSE	126
TABLE 3.36: ISSUES FOUND FOR THE TRANSLATION OF THE SUBCATEGORY OF PASSIVE VOICE	
PRESENT CONTINUOUS TENSE	127
TABLE 3.37: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF ACTIVE VOICE	
PRESENT CONTINUOUS TENSE	127
TABLE 3.38: ISSUES FOUND FOR THE TRANSLATION OF THE SUBCATEGORY OF ACTIVE VOICE	
PRESENT CONTINUOUS TENSE	128
TABLE 3.39: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF PASSIVE VOICE	
PRESENT CONTINUOUS TENSE	129
TABLE 3.40: ISSUES FOUND FOR THE TRANSLATION OF THE SUBCATEGORY OF PASSIVE VOICE	
PRESENT CONTINUOUS TENSE	129
TABLE 3.41: SUBCATEGORIES WITHIN THE CATEGORY OF ADVERBIALS	131
TABLE 3.42: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>BY</i>	132
TABLE 3.43: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>FOR</i>	133
TABLE 3.44: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>WHEN</i>	134
TABLE 3.45: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>BEFORE</i>	134
TABLE 3.46: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>AFTER</i>	135
TABLE 3.47: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>WITHOUT</i>	135
TABLE 3.48: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>WHILE</i>	136
TABLE 3.49: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY \emptyset	137
TABLE 3.50: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>BY</i>	137
TABLE 3.51: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>FOR</i>	137
TABLE 3.52: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>WHEN</i>	138
TABLE 3.53: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>BEFORE</i>	139
TABLE 3.54: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>AFTER</i>	139
TABLE 3.55: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>WITHOUT</i>	140
TABLE 3.56: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY <i>WHILE</i>	140
TABLE 3.57: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
PRECEDED BY \emptyset	141
TABLE 3.58: SUBCATEGORIES WITHIN THE CATEGORY OF REFERENTIALS	142
TABLE 3.59: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
FUNCTIONING AS NOUNS	142
TABLE 3.60: ISSUES FOUND FOR THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS	
FUNCTIONING AS NOUNS	143
TABLE 3.61: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
FUNCTIONING AS OBJECTS OF CATENATIVE VERBS	144
TABLE 3.62: FRENCH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS	
FUNCTIONING AS OBJECTS OF PREPOSITIONAL VERBS	145

TABLE 3.63: SPANISH TRANSLATION STRUCTURES FOR THE SUBCATEGORY OF -ING WORDS FUNCTIONING AS NOUNS	145
TABLE 3.64: ISSUES FOUND FOR THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS FUNCTIONING AS NOUNS	146
TABLE 3.65: ISSUES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS FUNCTIONING AS OBJECTS OF CATENATIVE VERBS	147
TABLE 3.66: ISSUES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS FUNCTIONING AS OBJECTS OF PREPOSITIONAL VERBS	148
TABLE 3.67: SUBCATEGORIES OF THE CATEGORY TITLES WITH INCORRECT EXAMPLES	149
TABLE 3.68: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS AT THE BEGINNING OF TITLE.....	150
TABLE 3.69: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS EMBEDDED WITHIN QUOTATION MARK EMBEDDED IN SENTENCES.....	151
TABLE 3.70: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF TITLES STARTING WITH -ING WORDS EMBEDDED WITHIN QUOTATION MARKS.....	151
TABLE 3.71: SUBCATEGORIES OF THE CATEGORY PROGRESSIVES WITH INCORRECT EXAMPLES	152
TABLE 3.72: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF ACTIVE VOICE PRESENT TENSE	152
TABLE 3.73: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF PASSIVE VOICE PRESENT TENSE	153
TABLE 3.74: SUBCATEGORIES OF THE CATEGORY ADVERBIALS WITH INCORRECT EXAMPLES..	154
TABLE 3.75: : ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF WHEN + ING	155
TABLE 3.76: : ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF FOR + ING	155
TABLE 3.77: SUBCATEGORIES OF THE CATEGORY CHARACTERISERS WITH INCORRECT EXAMPLES	156
TABLE 3.78: : ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF PRE-MODIFIERS.....	156
TABLE 3.79: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF REDUCED RELATIVE CLAUSES	157
TABLE 3.80: SUBCATEGORIES OF THE CATEGORY REFERENTIALS WITH INCORRECT EXAMPLES	157
TABLE 3.81: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF CATENATIVES	158
TABLE 3.82: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF GERUNDIAL	158
TABLE 3.83: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF PREPOSITIONAL VERBS	159
TABLE 3.84: SUBCATEGORIES OF THE CATEGORY TITLES WITH INCORRECT EXAMPLES	160
TABLE 3.85: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS AT THE BEGINNING OF INDEPENDENT TITLES	161
TABLE 3.86: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS AT THE BEGINNING OF TITLE WITHIN QUOTATION MARKS EMBEDDED IN SENTENCES	161
TABLE 3.87: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF -ING WORDS AT THE BEGINNING OF TITLES WITHIN QUOTATION MARKS AT THE BEGINNING OF SENTENCES.....	162
TABLE 3.88: SUBCATEGORIES OF THE CATEGORY CHARACTERISERS WITH INCORRECT EXAMPLES	162
TABLE 3.89: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF PRE-MODIFIERS.....	163
TABLE 3.90: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF REDUCED RELATIVE CLAUSES	163
TABLE 3.91: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF NOMINAL ADJUNCTS.....	163

TABLE 3.92: SUBCATEGORIES OF THE CATEGORY ADVERBIALS WITH INCORRECT EXAMPLES ..	164
TABLE 3.93: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF BY + ING ..	165
TABLE 3.94: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF WHEN + ING	165
TABLE 3.95: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF FOR + ING	166
TABLE 3.96: SUBCATEGORIES OF THE CATEGORY PROGRESSIVES WITH INCORRECT EXAMPLES	167
TABLE 3.97: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF ACTIVE VOICE PRESENT TENSE	167
TABLE 3.98: SUBCATEGORIES OF THE CATEGORY REFERENTIALS WITH INCORRECT EXAMPLES	168
TABLE 3.99: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF CATENATIVES	168
TABLE 3.100: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF GERUNDIAL NOUNS	169
TABLE 3.101: ERROR TYPES FOUND IN THE TRANSLATION OF THE SUBCATEGORY OF PREPOSITIONAL VERBS	169
TABLE 3.102: OVERALL AUTOMATIC METRIC SCORES FOR EACH TARGET LANGUAGE	172
TABLE 3.103: PEARSON R CORRELATION BETWEEN AUTOMATIC SCORES FOR SPANISH	173
TABLE 3.104: PEARSON R CORRELATION BETWEEN AUTOMATIC SCORES FOR FRENCH.....	173
TABLE 3.105: PEARSON R CORRELATION BETWEEN AUTOMATIC SCORES FOR GERMAN	173
TABLE 3.106: PEARSON R CORRELATION BETWEEN AUTOMATIC SCORES FOR JAPANESE.....	174
TABLE 3.107: PEARSON R CORRELATION SCORES BETWEEN AUTOMATIC METRICS AND TARGET LANGUAGES	174
TABLE 3.108: PEARSON R CORRELATION SCORES BETWEEN AUTOMATIC METRICS AND TARGET LANGUAGES CALCULATED BASED ON THE AVERAGES OBTAINED BY GROUPING THE EXAMPLES ACCORDING TO THE NUMBER OF EVALUATORS WHO CONSIDERED THEM CORRECT	178
TABLE 4.1: EXAMPLES WHERE <i>YOU</i> IS THE SUBJECT OF BOTH THE MAIN AND THE SUBORDINATE CLAUSE COMPLETED BY IMPERATIVES AND MODAL VERBAL PHRASES AS THE NUCLEI OF THE MAIN PREDICATE	187
TABLE 4.2: NUMBER OF RELEVANT AND IRRELEVANT EXAMPLES IN THE RULE DEVELOPMENT SETS TO MEASURE PRECISION AND RECALL	190
TABLE 4.3: RELEVANT EXAMPLES OF REDUCED RELATIVE CLAUSES MISSED BY THE RULE.....	191
TABLE 4.4: EXAMPLES OF RELEVANT -ING WORDS NOT RETRIEVED BY THE RULE	191
TABLE 4.5: EXAMPLES OF NON-RELEVANT SENTENCES IDENTIFIED BY THE RULE.....	192
TABLE 4.6: EXAMPLES OF DANGLING SUBJECTS RETRIEVED BY THE RULE.....	192
TABLE 4.7: EXAMPLES OF DANGLING IMPLICIT SUBJECTS NOT RETRIEVED BY THE RULE.....	193
TABLE 4.8: EXAMPLES WHICH ARE NOT DANGLING SUBJECTS RETRIEVED BY THE RULE	193
TABLE 4.9: EXAMPLES OF MISSING ARTICLES IN THE CONTEXT OF -ING WORDS RETRIEVED BY THE RULE	195
TABLE 4.10: EXAMPLES OF MISSING ARTICLES NOT RETRIEVED BY THE RULE	195
TABLE 4.11: EXAMPLES WHICH ARE NOT MISSING ARTICLES RETRIEVED BY THE RULE	195
TABLE 4.12: TRANSFORMATION: -ING INTO NOUN + OF	199
TABLE 4.13: EXAMPLES OF TRANSFORMATIONS	200
TABLE 4.14: TRANSFORMATION: -ING INTO NOUN + OF	201
TABLE 4.15: TRANSFORMATION FOR FRENCH, GERMAN AND JAPANESE	203
TABLE 4.16: ORIGINAL DATA FOR JAPANESE	203
TABLE 4.17: TRANSFORMATION OF PROGRESSIVE ASPECT INTO SIMPLE TENSE FOR JAPANESE .	204
TABLE 4.18: TRANSFORMATION OF PROGRESSIVE ASPECT INTO SEVERAL DESIDERATIVE STRUCTURES	204
TABLE 4.19: TRAINING MATERIAL FOR MOSES	218
TABLE 4.20: SUMMARY OF CHARACTERISTICS PER IMPROVEMENT APPROACH	220

TABLE 5.1: RESULTS OBTAINED IN THE HUMAN EVALUATION FOR THE EXAMPLES ADDRESSED BY THE NEW CL RULES	223
TABLE 5.2: EXAMPLES OF REWRITES FOR THE NEW CL RULES.....	224
TABLE 5.3: MAXIMUM MARGIN FOR IMPROVEMENT FOR THE CL RULES	225
TABLE 5.4: EVALUATION RESULTS FOR CL-RULE1	226
TABLE 5.5: EVALUATION RESULTS FOR CL-RULE2	226
TABLE 5.6: EVALUATION RESULTS FOR CL-RULE3	226
TABLE 5.7: RESULTS FOR RULE TRANSFORMING -ING TITLES INTO IMPERATIVES FOR SPANISH	232
TABLE 5.8: RESULTS FOR RULE TRANSFORMING -ING TITLES INTO NOUNS FOR SPANISH	232
TABLE 5.9: RESULTS FOR RULE TRANSFORMING -ING TITLES INTO NOUNS FOR FRENCH.....	232
TABLE 5.10: RESULTS FOR RULE TRANSFORMING -ING ADVERBIALS INTRODUCED BY <i>WHEN</i> FOR FRENCH.....	233
TABLE 5.11: RESULTS FOR RULE TRANSFORMING -ING TITLES INTO NOUNS FOR GERMAN	233
TABLE 5.12: RESULTS FOR RULE TRANSFORMING -ING TITLES INTO NOUNS FOR JAPANESE.....	233
TABLE 5.13: RESULTS FOR RULE TRANSFORMING -ING ADVERBIALS INTRODUCED BY <i>WHEN</i> FOR JAPANESE.....	234
TABLE 5.14: RESULTS FOR THE RULE REMOVING ARTICLES IN FRONT OF INFINITIVE VERBS AT THE BEGINNING OF TITLES FOR SPANISH	235
TABLE 5.15: RESULTS FOR THE RULE TRANSFORMING THE WORD CLASS OF THE OBJECTS OF CATENATIVE VERBS FOR SPANISH	236
TABLE 5.16: RESULTS FOR THE RULE TRANSFORMING THE WORD CLASS OF THE OBJECTS OF CATENATIVE VERBS FOR FRENCH.....	236
TABLE 5.17: RESULTS FOR SENTENCE-LEVEL QUALITY FOR SPANISH	238
TABLE 5.18: RESULTS FOR -ING WORDS-LEVEL QUALITY FOR SPANISH	238
TABLE 5.19: CROSS-TABULATION OF -ING WORD AND SENTENCE-LEVEL RESULTS	238
TABLE 5.20: RESULTS FOR SENTENCE-LEVEL QUALITY FOR FRENCH.....	239
TABLE 5.21: RESULTS FOR -ING WORD-LEVEL QUALITY FOR FRENCH	239
TABLE 5.22: CROSS-TABULATION OF -ING WORD AND SENTENCE-LEVEL RESULTS	239
TABLE 5.23: RESULTS FOR SENTENCE-LEVEL QUALITY FOR GERMAN	240
TABLE 5.24: RESULTS FOR -ING WORD-LEVEL QUALITY FOR GERMAN	240
TABLE 5.25: CROSS-TABULATION OF -ING WORD AND SENTENCE-LEVEL RESULTS	240
TABLE 5.26: OVERALL IMPROVEMENT RESULTS FOR THE TESTED -ING WORDS	242

LIST OF FIGURES

FIGURE 1.1: RELATIONSHIP BETWEEN THE 3 RULE-BASED APPROACHES REPRESENTED USING VAUQUOIS TRIANGLE (FROM ARNOLD ET AL. 1994: 77)	15
FIGURE 2.1: MODEL OF THE RULE USED TO EXTRACT REDUCED RELATIVE CLAUSES.	63
FIGURE 2.2: EQUATION FOR THE KAPPA COEFFICIENT, WHERE P_{OBS} IS THE OBSERVED AGREEMENT AND P_{EXP} IS THE EXPECTED AGREEMENT.	79
FIGURE 2.3: FORMULA TO CALCULATE SAMPLE SIZE (SS), WHERE Z IS THE CONFIDENCE LEVEL VALUE (E.G. 1.96 FOR 95% CONFIDENCE LEVEL), P IS THE DEGREE OF VARIABILITY (0.5 USED, MAXIMUM VARIABILITY), AND C IS THE CONFIDENCE INTERVAL, EXPRESSED AS A DECIMAL (E.G. $0.04 = \pm 4$).	82
FIGURE 2.4: EQUATION FOR BLEU, WHERE THE MODIFIED N-GRAM PRECISION P_N IS CALCULATED FOR N-GRAMS UP TO NUMBER N AND USING POSITIVE WEIGHTS w_N AND IT IS MULTIPLIED BY THE EXPONENTIAL BREVITY PENALTY FACTOR BP WHERE C IS THE LENGTH OF THE MT OUTPUT AND R IS THE LENGTH OF THE REFERENCE TEXT TRANSLATION.	90
FIGURE 2.5: EQUATION FOR NIST, WHERE B IS THE BREVITY PENALTY FACTOR $= 0.5$ WHEN THE NUMBER OF WORDS IN THE SYSTEM OUTPUT IS $2/3$ OF THE AVERAGE NUMBER OF WORDS IN THE REFERENCE TRANSLATION, N IS 5 AND L_{REF} IS THE AVERAGE LENGTH OF THE REFERENCE TRANSLATIONS AND L_{SYS} IS THE LENGTH OF THE MT TRANSLATION.	91
FIGURE 2.6: EQUATION FOR GTM, WHERE THE FMEASURE IS THE COMPOSITE OF P (PRECISION) AND R (RECALL) AND WHERE THE INTERSECTION OF THE ELEMENTS CT (CANDIDATE TEXT) AND RT (LENGTH OF THE REFERENCE TEXT) ARE COMPUTED USING AN EXTENDED FORM OF THE MMS (MAXIMUM MATCHING SIZE) WHICH REWARDS GROUPS OF LONGER ADJACENT MATCHING WORDS R (RUN).	92
FIGURE 2.7: EQUATION FOR METEOR, WHERE THE FMEAN IS THE COMBINATION OF P (PRECISION) AND R (RECALL) AND A MAXIMUM PENALTY OF 0.5 IS INTRODUCED TO ACCOUNT FOR NON-ADJACENT UNIGRAMS.	93
FIGURE 2.8: EQUATION FOR TER, WHERE THE NUMBER OF EDITS IS DIVIDED BY THE AVERAGE NUMBER OF WORDS IN THE REFERENCE.	94
FIGURE 3.1: COMPARISON BETWEEN NORMALISED AUTOMATIC SCORES AVERAGED FOR THE NUMBER OF EVALUATORS WHO CONSIDERED THE -ING WORDS CORRECT FOR SPANISH ..	176
FIGURE 3.2: COMPARISON BETWEEN NORMALISED AUTOMATIC SCORES AVERAGED FOR THE NUMBER OF EVALUATORS WHO CONSIDERED THE -ING WORDS CORRECT FOR FRENCH ...	176
FIGURE 3.3: COMPARISON BETWEEN NORMALISED AUTOMATIC SCORES AVERAGED FOR THE NUMBER OF EVALUATORS WHO CONSIDERED THE -ING WORDS CORRECT FOR JAPANESE	177
FIGURE 3.4: COMPARISON BETWEEN NORMALISED AUTOMATIC SCORES AVERAGED FOR THE NUMBER OF EVALUATORS WHO CONSIDERED THE -ING WORDS CORRECT FOR GERMAN..	177
FIGURE 4.1: MODEL OF A SIMPLE RULE FOR ACROLINX IQ	185

LIST OF ABBREVIATIONS

AECMA	Association Européenne des Constructeurs de Matériel Aérospatial
AdjP	Adjectival Phrase
CFE	Caterpillar Fundamental English
CL	Controlled Language
CLAW	Controlled Language Application Workshop
CTE	Caterpillar Technical English
EAGLES	Expert Advisory Group on Language Engineering Standards
EBMT	Example-Based Machine Translation
FEMTI	Framework for the Evaluation of Machine Translation in ISLE
GTM	General Text Matcher
ILSAM	International Language for Service and Maintenance
ISLE	International Standards for Language Engineering
IT	Information Technology
JJ	Adjective
MT	Machine Translation
NN	Noun
NP	Noun Phrase
PAHO	Pan-American Health Organization
PEP	Plain English Program
PE	Post-editing
POS	Part Of Speech
PP	Prepositional Phrase
QA	Quality Assurance
RBMT	Rule-Based Machine Translation
Regex	Regular Expressions
SE	Simplified English
SL	Source Language
SMT	Statistical Machine Translation
SPE	Statistical Post-editing
S&R	Search and Replace

TCI	Translation Confidence Index
TL	Target Language
TM	Translation Memory
STS	Systran Translation Stylesheet
UD	User Dictionary
VBG	Gerund
XML	eXtensible Markup Language

-ING WORDS IN RBMT: MULTILINGUAL EVALUATION AND EXPLORATION OF PRE- AND POST-PROCESSING SOLUTIONS

Nora Aranberri Monasterio
School of Applied Language and Intercultural Studies
Dublin City University

Abstract

This PhD dissertation falls within the domain of machine translation and it specifically focuses on the machine translation of IT-domain -ing words into four target languages: French, German, Japanese and Spanish. Claimed to be problematic due to their linguistic flexibility, i.e. -ing words can function as nouns, adjectives and verbs, this dissertation investigates how problematic -ing words are and explores possible solutions for improvement of their MT output.

A corpus-based approach for a better representation of the domain-specific structures where -ing words occur is used. After selecting a significant sample, the -ing words are classified following a functional categorisation presented by Izquierdo (2006). The sample is machine-translated using a customised RBMT system.

A feature-based human evaluation is then performed in order to obtain information about the specific feature under study. The results showed that 73% of the -ing words were correctly translated in terms of grammaticality and accuracy for German, Japanese and Spanish. The percentage for French was lower at 52%. These data, combined with a thorough analysis of the MT output, allows for the identification of cross-language and language-specific issues and their characteristics, setting the path for improvement.

The approaches for improvements examined cover both the pre- and post-processing stages of automated translation. For pre-processing, controlled language (CL) and automatic source re-writing (ASR) are explored and evaluated. For post-processing, global search and replace (Global S&R) and statistical post-editing (SPE) methods are tested. CL is reported to reduce -ing word ambiguity but to not achieve substantial machine translation improvement. Regex-based implementations of ASR and Global S&R efforts show considerable translation improvements ranging from 60% to 95% and minimal degradation, ranging from 0% to 18%. The results yielded for SPE show little improvement, or even degradation at both sentence and -ing word level.

INTRODUCTION

This dissertation arises from a need to better understand issues that effect the production of higher quality machine translation (MT) output. It examines the performance of rule-based machine translation (RBMT) in respect of -ing words and further investigates remediation techniques designed to improve MT system performance.

The research idea stemmed from discussion between industry partners, Symantec, and researchers in controlled language (CL), machine translation and post-editing (PE) at Dublin City University (DCU). The need to better understand the impediments to better quality MT led to a successful application for research funding to Enterprise Ireland.¹ The core objective of this application was to improve the process of machine translation in an industrial context, in particular, that of Symantec Ltd. IT-domain User Documentation.

RESEARCH OBJECTIVES

BACKGROUND RESEARCH

In the early days of post-editing in the IT vertical in 2006, vendors reported problems with MT output which they deemed intractable, that is, they claimed that it was often easier to translate certain sentences or linguistics features from scratch than to try to fix the MT output. The handling of software strings, which required special treatment, -ing words, passive structures and long sentences were the top issues for Symantec at the start of this dissertation. While the manipulation of special strings was being studied (Roturier and Lehmann, 2009), the grammatical features remained to be addressed. According to the company's Style Guide, passive structures should be avoided and sentence length restricted to 25 words. Compliance with the Style Guide was enforced with CL rules in the authoring stage. Symantec was particularly concerned about the MT performance of -ing words because the Style Guide allowed their use and technical

¹ This research was funded by a joint Enterprise Ireland/Symantec Innovation Partnerships Fund (IP-2006-0368) for the first two years and funded completely by Symantec Ltd., Dublin, for the third year.

Enterprise Ireland: www.enterprise-ireland.com
Symantec Ltd.: www.symantec.com

writers were not content with the use of the CL checker to address the issue. They claimed that -ing words had various uses and could not be removed from content. In addition, little empirical data or analytic assessment was available for the problem. As MT sits between the authoring stage and the PE stage and both parties were reluctant to address the issue, Symantec felt impelled to investigate the nature of -ing words and their MT performance to direct appropriate action in both stages.

Previous related research (Roturier, 2006) had already pointed at -ing words as a prominent problem. Roturier aimed at establishing a CL in the Symantec document authoring stage and identified a number of rules to be deployed, including the general “avoid -ing words” rule. This rule had a double objective: first, to *“remove the ambiguity created by specific -ing words to improve the readability of source XML topics”* and secondly, to *“avoid certain -ing words to improve the performance of an RBMT system”* (Aranberri and Roturier, 2009: 1). With the assumption that -ing words in general were problematic, the rule focused on maximising the number of identified cases and not much effort was put into the rate of precision, i.e. not flagging the -ing words which would be handled correctly by the MT system. As a result, many examples were identified as errors by the CL checker even though they were acceptable for MT. This proved confusing for writers, as well as time-consuming, as both problematic and non-problematic -ing words needed attention. The identification of unwanted instances has an added effect: the identified unproblematic sentences, which obviously remain unchanged, add to the unresolved number of problems, unfairly affecting the final automatic scoring of problems in a particular document.

In his dissertation Roturier (2006) pointed out the existence of different contexts for -ing words, which were not all equally problematic. However, due to the high number of CL rules to be evaluated (54) in his research, the examples included for each rule were low, including only 31 for -ing words, insufficient to get an insight into this particular feature. From a more theoretical perspective, scholars have argued that -ing words are problematic for both humans and MT due to their grammatical flexibility and have listed them within the top problematic issues in works focusing on machine translation or in the Global English Style Guide (Bernth and McCord, 2000; Bernth and Gdaniec, 2001; Kohl, 2008). Authoring and translation technology-related applications are concerned with the use of -ing words (Adriaens and Schreurs, 1992;

Wells Akis and Sisson, 2002; O'Brien 2003). Yet, to the best knowledge of the author, no empirical data has been provided to confirm the claim that -ing words are among the most problematic linguistic features for English as a source language.

RESEARCH QUESTIONS

This led to the first main research question of this dissertation: what are the most problematic -ing words for RBMT systems when translating IT user guides into French, German, Japanese and Spanish?

Once the problematic -ing words were identified, the second main research question could be investigated, i.e. what approaches were most efficient for the improvement of the machine translation quality of -ing words. We hoped to answer this question by testing different techniques for a number of problematic subcategories of -ing words, along with the working details of each technique, their advantages and weaknesses, effectiveness on MT quality and possibilities for implementation.

INDUSTRY-ACADEMIA RESEARCH COLLABORATION

The funding for this research was awarded through the Innovation Partnerships Programme of Enterprise Ireland, the government agency in charge of promoting and developing national business. This programme aims at supporting joint industry and academic efforts to pursue commercially beneficial research. Industry-academia collaborations have been mainly short-term and *ad hoc* initiatives, which are only recently receiving attention from governments and supranational institutions such as the European Union (CREST, 2008).

Research objectives in academia and industry R&D teams are different. As Kenney-Wallace (2001) describes, industry wants results, cutting edge research. The research tends to be task-specific, performed under market direction and time constraints. Academia, conversely, seeks academic excellence and it often fails on applicability (ibid). An added difference between the two research strategies lies on the fact that academic research is public whereas industrial research usually has private aspirations, with great emphasis on intellectual property (IP) ownership (Carpenter et al. 2004). When combined, industry exposes complex real-world scenarios and problems to be solved to academics, and offers access to world-class facilities, resources and industry

experts, as well as providing financial support. Academia, in turn, contributes with a solid knowledge of existing research work and competing approaches, and scientific rigour, logical structures and frameworks so that research results are valid and hold for similar scenarios, enabling replication.

Collaboration involves some challenges. Among them is the need to adapt to market-bound projects. The pressure is not only a consequence of the speed to market, but also of the applicability within the market. Concepts and products evolve and the research must follow. Similarly, companies depend on the market and the collaborations risk coming to an abrupt termination if they do not “*fit with the [company’s] new strategic plan*” (Carpenter et al. 2004: 1003). The availability of resources is also a double-edged sword. On the one hand, it makes research possible and greatly broadens the possibilities and scope. On the other hand, depending on them can lead to important gaps once the collaboration is terminated (Carpenter et al. 2004). Also, companies contribute with substantial funding and resources but they are not limitless and flawless.

This dissertation is a product of industry-academia collaboration. It exemplifies the advantages facilitated by an industry-academia setting – access to cutting-edge MT, authoring technology, and industrial experts, financial support to perform human evaluations, linguistic support for multilingual research – while also presenting limitations – budgetary constraints for human evaluations, availability of identical resources across languages for statistical training. But ultimately, with the right infrastructure in place and the awareness of the different research drivers in industry and academia, a joint venture has the potential to achieve rigorous cutting edge research with direct implementation. Taking both the advantages and disadvantages of this kind of collaboration on board, we hope to have attained usefull industrial process modifications based on the latest scientific principles.

SYMANTEC

Symantec was founded in 1982 and is now one of the world's leading software companies with more than 17,500 employees in more than 40 countries. As a global IT company, it provides security, storage and systems management solutions for individual consumers to small businesses and large enterprises. Whereas the Global

Headquarters are based in Cupertino, CA, global operations are divided into three regions: Americas, EMEA (Europe, Middle East and Africa) and APJ (Asia Pacific and Japan).

The EMEA localisation headquarters are based in Dublin. This group is responsible for the translation and adaptation of software and deliverables for the locales in their region. This involves the translation of tens of millions of words per year, as well as Software QA Testing. With the need for faster turnaround times in order to achieve *simship*, that is, the simultaneous shipment of all localised versions, the localisation department has a strong focus on innovation in translation technology.

SYMANTEC LOCALISATION WORKFLOW

The first technology introduced into the human translation workflow was that of translation memories (TM), dating back to 1997. This allowed for significant reuse of existing translations and increased consistency. Research on the integration of machine translation (MT) technology started in 2003 (Roturier, 2009). The desktop applications of TM and MT led to the initial use of MT for Technical Support translation in 2005. In 2006 the server-based enterprise system was installed, with user documentation for EMEA and APJ benefiting from it from 2006 and 2008 respectively, thanks to the deployment and support offered by the internal Symantec team.

MT technology is used in conjunction with TM technology. All translation segments with 85% fuzzy match or above are submitted for human translation whereas segments falling below that threshold are machine translated and sent for post-editing (ibid). The first enterprise product localised with deployment of the MT technology was completed in 7 days, as opposed to the 15 days (best performance) it took to localise a language version of the same version of the same product without MT. Not only did turnaround time decrease, consistency is also said to have improved (ibid).

The MT system used at Symantec is SYSTRAN. Although it is a proprietary system, it offers a number of customisation techniques which help contextualise the system and improve its translation quality. Symantec exploits two main techniques: user dictionaries (UD) and translation stylesheets (STS). Project-specific or domain-specific UD's are created and re-used, which contain the translation preferred by Symantec. STS procure context-sensitive translations. Based on tags, certain parts of the texts, such as

Graphical User Interface (GUI) options are set to be translated separately, or strings of commands set as “do not translate”.

The MT system requirements pervade the translation process infrastructure. As with most large companies, Symantec has developed its own style guide for document authoring. However, ensuring all writers comply with the guide is not straightforward. In order to achieve this, a Controlled Language checker (acrolinx™ IQ suite) was introduced in 2005, which is used by the writing teams during the document authoring process. Compliance with spelling, terminology, grammar and style rules is guaranteed and made easier. Roturier (2009) showed that the lower the number of CL rule violations, the better the MT quality. In addition to ensuring the quality and consistency of source documents, the use of a CL checker also offers the possibility of producing more suitable documents for MT, as MT-specific rules can be added to the style guide recommendations.

At the final stage of the translation workflow, Symantec established a post-processing module based on global search and replace rules applied through regular expressions (Roturier et al. 2005). This module aims at automatically fixing repetitive errors that cannot be fixed with other SYSTRAN customisation possibilities, before the output is passed on to human post-editors.

RELEVANCE OF SYMANTEC-BASED RESEARCH

As this research was a collaboration with Symantec, it is only logical that the research objectives are somewhat rooted within their particular workflow. This means that the IT documents analysed pertain to Symantec products, that the MT system used is the rule-based SYSTRAN and that the target languages examined – French, German, Japanese and Spanish – are the ones of interest to the company. However, a look into the reports of large IT companies and the theoretical questions involved demonstrates that the research is relevant to the field in general.

As discussed in Chapter 2 section 2.1.1, the grammatical characteristics of a particular text type are normally consistent across documents. Therefore, by considering corpus-design issues to ensure a balanced and representative corpus, the generalisability of the results to IT-domain procedural and descriptive texts is valid. Regarding the use of SYSTRAN, whereas we acknowledge the use of a single MT

system, it could be argued that since SYSTRAN is a widely used system, other companies could benefit from the research.² In addition, although the overall results might depend on the development and customisation level of SYSTRAN at Symantec, specific results might hold for RBMT systems in general – ProMT, PAHOMTS, SDL KbTS. Moreover, as described in Chapter 1 section 1.2.2 hybrid models are emerging and even statistical machine translation (SMT) architectures are developing towards the inclusion of linguistic knowledge in their systems. Current SMT providers, such as Language Weaver or Microsoft, are reporting the introduction of syntactic and semantic information into their systems (Wendt, 2008; Language Weaver News, 2008). The move towards hybrid MT systems underlines the importance of research on linguistically-based architectures and techniques represented in this research. The choice of target languages was related to the commercial importance they hold within Symantec (among other characteristics which are discussed in Chapter 2 section 2.2.1.5). According to the ratings of World Online Wallet (WOW), Japanese, German, Spanish and French are, after English, the languages which bring the biggest benefits, hence their importance (DePalma and Kelly, 2009). As a result, large IT companies and language providers alike – Microsoft, IBM, SDL – include the selected languages within their target language range. In conclusion, it can be argued that despite its origin in Symantec, the results obtained from the research can be of benefit to a variety of players within the localisation industry.

STRUCTURE OF THE DISSERTATION

Chapter 1 provides an overview of the most relevant literature for this research, discussing notions such as -ing words, machine translation, evaluation and controlled language, and the research that has been conducted to date.

The initial investigation of “gerunds” as a problem for MT led to the rapid realisation that “gerund” was in fact not an appropriate term for describing the linguistic feature we wished to study because it restricted the structures and types of -ing words it covered, but in fact the more general term “-ing word” was more

² Companies and institutions which work with SYSTRAN include, among others, CISCO, Daimler-Chrysler, EADS and the European Union.

appropriate.³ This is still a very broad category and, as Chapter 1 and, more specifically, Chapter 2 outlines, a much more detailed categorisation was necessary in order to proceed with the quantification and evaluation of the problem.

Chapter 2 then focuses on the methodologies used to build a balanced corpus of IT documents. The functional -ing word classification is introduced and the approach taken to extract the -ing words pertaining to each category described. The rationale and preparations for the human and automatic evaluations are examined as well as setting the path for the analysis.

Chapter 3 presents the evaluation results obtained from the human judges and automatic metrics. This gives an insight into the handling of -ing word subcategories by the RMBT system and accounts for the specific problems it faces. This human evaluation is then compared against commonly used automatic metrics to show the usefulness of the metrics for a feature-based evaluation.

Chapter 4 reviews the localisation workflow identifying techniques for translation quality improvement in the pre- and post-processing stages. The procedure for implementation of four techniques – Controlled Language, Automatic Source Re-writing, Global Search & Replace and Statistical Post-editing – is described by applying them for a number of problematic subcategories of -ing words. It particularly focuses on the details in creating effective rules for rule-based techniques, and measuring their precision and recall, as well as the considerations for data-driven techniques.

Chapter 5 reports the translation improvement obtained from the implementation of the techniques described in Chapter 4 for a number of -ing word subcategories. It describes the human evaluation performed for each technique and reports the results, as well as considering possibilities for deployment within the localisation workflow.

Finally, Chapter 6 summarises the findings of this dissertation and reviews and critiques the methodologies used. Additionally, it also discusses the implications of performing collaborative research between industry and academia. Finally, it identifies directions for future research.

³ The definition of the terminology is given in Chapter 1 section 1.1.

CHAPTER 1

CHAPTER 1: LITERATURE REVIEW

In this Chapter we review the different domains within the framework of machine translation quality improvement. We start by defining the object of research, i.e. -ing words and by discussing the lack of agreement on their classification by scholars. We then introduce machine translation (MT) and its architectures, focusing on rule-based systems. In section 3 we proceed to describe the processing difficulties -ing words pose for both humans and computers.

Section 4 focuses on MT evaluation. We first consider the traditional human evaluation. We touch upon the benefits of human knowledge to perform this task and the challenges relating to subjectivity. We then turn to the recent approach of automatic evaluation to present its principles, advantages in terms of time and cost, and pitfalls in terms of informativeness.

Finally, section 5 examines the possibilities which are currently researched and are in use in real-life workflows for improving MT output quality. We focus on techniques which cover the pre-MT and post-MT stages: Controlled Language (CL), and automatic post-editing (APE).

1.1 DEFINING -ING WORDS

Traditional grammars (Quirk et al. 1985) divide words with the -ing suffix into gerunds and present participles. Gerunds are described as deverbal forms which function as nouns (a). Present participles, in turn, are divided into two subgroups. Firstly, deverbal forms that function as adjectives (b). Secondly, forms which, in combination with auxiliaries, constitute the progressive aspect (c).

(a) *Allocating fewer CPU cycles to a backup job may result in slower backup performance.*

(b) *The program will use the existing password.*

(c) *You must have administrative rights of the computer you are using.*

However, a number of current grammarians claim that this division can no longer be defended. According to Huddleston and Pullum (2002: 82-83):

Historically the gerund and present participle of traditional grammar have different sources, but in Modern English the forms are identical. [...] The historical difference is of no relevance to the analysis of the current

inflectional system. [...] We have therefore just one inflectional form of the verb marked by the -ing suffix; we label it with the compound term 'gerund-participle' for the verb-form, as there is no reason to give priority to one or other of the traditional terms.

These authors sustain their claim by rejecting any difference in form, function and aspect between the traditional gerund and present participle. In reference to form, they acknowledge an internal difference. Gerunds allow for the subject to take genitive cases and use plain or accusative cases for more informal uses (d). The present participles, however, do not allow for genitives and restrict the use of accusatives to informal style (e). However, this difference only accounts for the instances where the -ing clause allows a subject. Therefore, it is not an absolute difference that can always be applied as a specific characteristic of all gerunds and present participles.

(d) *She resented his/him/*he being invited to open the debate.*⁴

(e) *We appointed Max, he/him/*his being much the best qualified of the candidates.*

In reference to function, they claim that the traditional distinction between gerunds and present participles is based on the former acting as nouns whereas the latter act as adjectives. They reject the traditional practice of using the part-of-speech function as the criterion to classify clauses, but they prove that, even by using this method, it is not possible to account for the distinction between gerunds and present participles. They use the example of catenative verbs, whose complements are classified as gerunds or present participles depending on their function as noun phrases (objects) or adjectival phrases (predicatives). They show that this distinction is not absolute because not all verbs that allow adjectival predicatives take gerund-participials (f), and not all present participles can be substituted by predicatives (g). Hence the claim that gerunds and present participles are not functionally nouns or adjectives but a form of the verb. In addition, they claim that the object/predicative distinction is also applied to infinitivals but in their case, no further distinction is sought. This gives ground to the belief that the distinction between gerunds and present participles is drawn from historical usage only.

(f) *They seemed resentful.*

**They seemed resenting it.*

(g) *He stopped staring at them.*

**He stopped calm.*

⁴ Examples d-i taken from Huddleston and Pullum (2002).

Finally, in reference to aspect, Huddleston and Pullum suggest that present participles occur as both modifiers and in combination with auxiliaries to form the progressive aspect. But one can claim that present participles do not always have a progressive meaning (h-i). Therefore, a clear distinction between gerunds, which are not supposed to show progressive aspect, and present participles cannot be drawn based on aspect either.

(h) *Hearing his cry, she dashed into the garden.*

(i) *Although having no TV himself, he was able to see the programme.*

In conclusion, it could be said that there is no systematic correlation of differences in form, function and aspect between the traditional gerund and present participle. As a result, Huddleston and Pullum (ibid) propose that words with a verb base and the -ing suffix be classified as gerundial nouns (genuine nouns), gerund-participles (forms of verbs) and participial adjectives (genuine adjectives).

The authors acknowledge difficulties in determining the boundaries between the groups, particularly when the -ing words appear on their own, with no further complementation. Still, they provide a set of clues to facilitate the identification process (see Table 1.1).

GERUNDIAL NOUNS	GERUND-PARTICIPLES	PARTICIPIAL ADJECTIVES
can take an “of” PP complement	can take NP objects	normally do not take NP objects
characteristically modified by adjectives	modified by adverbs	can be modified by degree adverbs such as “very” or “too”
combine with determiners	do not combine with determiners	verbs like “seem” take AdjP as complement
can very often inflect for plural	cannot inflect for plural	
can take genitives	cannot be modified by degree adverbs such as “very” or “too”	
	verbs like “seem” do not take AdjP as complement	

Table 1.1 Clues for classifying -ing words

The linguistic feature we aim at addressing is -ing words. These words acquire meaning in context. As we will describe in Chapter 2 section 2.1.3, the need to further categorise the contexts in which -ing words occur emerged to identify their purpose

and their accurate translation into different target languages. For instance, how should the -ing word *checking* be translated into Spanish? *Seleccionar* or *selección* or *seleccionando* or *de selección*? It is necessary to know the context in which the word appears to decide on the appropriate translation. This led us to the use of a functional classification scheme (Izquierdo, 2006), which uses the complementation of syntactic and semantic information to define this context. The context is delimited according to constituents, that is, groups of words that “*may behave as a single unit or phrase*” (Jurafsky and Martin, 2009: 419). In computational linguistics, constituents are identified following a set of rules which express “*the ways that symbols of the language can be grouped and ordered together*” (ibid: 421) and are represented by non-terminal symbols in parse trees, where their dependency is also apparent. This provides us with the syntactic group for which the -ing word is the head. We add a functional layer based on the premises of Functional Grammar (Halliday, 2004), which describes the purpose of the constituent, to the constituents and obtain the -ing functional constituent. For example, if we take the sentence (j) below, we see that *checking* is the head of the constituent *by checking the box on the left*, and this is an adverbial phrase of mode.

(j) *Accept this option by checking the box on the left.*

However, note that our interest is not that of evaluating entire -ing functional constituents, but the -ing words within them (see section 2.2.1.2 for a discussion on -ing word translation delimitation). In the example above, we are interested in whether *checking* is correctly translated given its context, that is, the functional constituent in which it occurs. We are not interested in whether *by* or *the box on the left* are correctly translated. In this dissertation, therefore, we use the term “-ing word” in reference to its functional constituent.

1.2 MACHINE TRANSLATION SYSTEMS

Let us briefly describe the architecture of MT systems before we consider the difficulties -ing words pose for them. Currently, MT systems, that is, “*computerised systems responsible for the production of translations from one natural language into another, with or without human assistance*” (Hutchins and Somers, 1992: 3) can be divided into rule-based or data-driven systems. The main difference lies in the method

used to acquire translation knowledge. Rule-based machine translation (RBMT) systems consist of bilingual dictionaries and thousands of grammatical rules governing translation from a specific source language (SL) to a specific target language (TL) manually encoded in their core modules. Data-driven systems, proposed as an alternative to circumvent the manual crafting of grammatical rules (Carl and Way 2003: xviii), extract all the necessary information for translation automatically from a bilingual parallel corpus used for “training”.

We find two main architectures within this paradigm: Statistical MT systems (SMT) and Example-based MT systems (EBMT). The basic principle behind SMT is that resources for translation can be extracted from corpora using statistical probabilities of distribution and estimation calculated from surface forms or words. The idea is to find the best translation probability given a source probability following Bayes’ theorem (Brown et al. 1990). The idea behind EBMT systems is to store a parallel corpus of already existing translations and then recombine fragments to produce new translations. We can distinguish three different stages in these systems: the matching of source segments against a bilingual translation database, the identification of corresponding translation segments, and the recombination of these to create the translation output (Nagao, 1984: 178).

In the following section, we describe the RBMT architecture in detail, as the MT system used for this research pertains to this category. We will revisit the SMT architecture when discussing post-processing approaches to improve MT output quality in section 1.5.2.2.

1.2.1 RULE-BASED MACHINE TRANSLATION

RBMT systems evolved from a naïve transformer architecture to an ambitious interlingua architecture, to end up with a more realistic transfer-based architecture, which is used nowadays. First to be proposed, transformer architectures used a bilingual dictionary to replace the words in the SL with the equivalents in the TL, minimising the amount of syntactic analysis (Arnold et al. 1994). This type of architecture proved too simplistic, assuming that languages could be translated word for word. The interlingua architectures aimed at creating a language-independent abstract representation of the SL which would then be used to recreate the TL (ibid).

No language-specific rules to relate a SL to a TL should be required. Such a method would allow for the introduction of different language pairs just by adding an analysis and a generation module. However, obtaining a deep abstract representation, based on the idea that all languages share primitive words to which the SL can be abstracted, proved to be very challenging. As a result, a third approach which included a degree of abstraction of the source and transfer rules and a generation module for the TL was devised (ibid). Although more time-consuming and costly due to the necessity of transfer rules for every language pair and the linguistic study required to create such rules, this approach proved more viable and accurate than the interlingual approach (for a graph on the relation between the three approaches, see Figure 1.1 below).

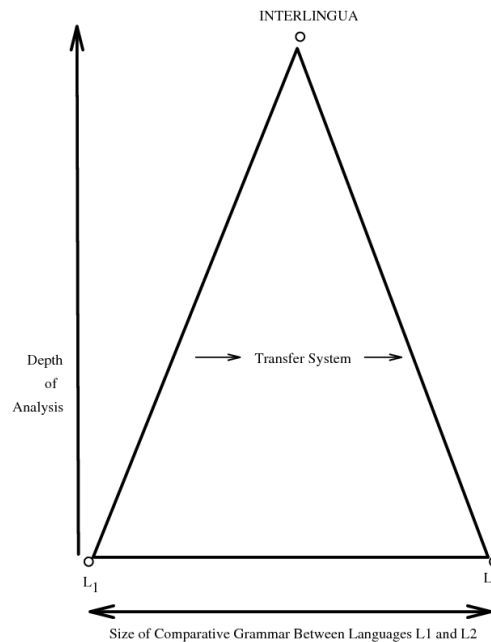


Figure 1.1: Relationship between the 3 rule-based approaches represented using Vauquouis Triangle (from Arnold et al. 1994: 77)

It is transfer architectures that have survived as commercial RBMT systems. SYSTRAN is an example of such a system and is the MT system in use in this study. Similar to other transfer systems, it consists of 3 modules: analysis, transfer, and generation. In the particular case of SYSTRAN, the generation module is divided into synthesis and rearrangement (Surcin et al. 2007). The former is performed using TL information only, the latter using both SL and TL information.

RBMT systems carry out a first cycle where the source text is analysed. SYSTRAN developers report that, for their system, 80% of the code belongs to the analysis module, whereas transfer accounts for 10% and synthesis and rearrangement would take up to 5% each (ibid). This clearly demonstrates the importance assigned to a high quality source analysis in the quest for a good translation.

SYSTRAN's analysis stage performs a Global Document Analysis, where the subject domain is identified and a grammatical analysis is performed (Senellart, 2007). The grammatical analysis extracts information about part-of-speech, clause dependencies and relationships between entities of the sentence as well as their functions (ibid). Several modules are involved in this process, which complete a representation of a source segment:

- Sentence segmentation
- Normalization of the languages
- Morphological analysis
- Grammatical disambiguation
- Clause identification
- Basic local relationships
- Enumeration analysis
- Predicate/Subject analysis
- Preposition Rattachement⁵
- Semantic Analysis

The transfer stage contains rules to transform SL structures and lexicons into TL structures and lexicons. This stage is dependent on the language pair. Finally, the generation stage performs the necessary synthesis (insertion of elements not explicitly present in the TL) and word-order rearrangement to model a grammatical TL output.

Bilingual dictionaries are paramount to this architecture. In an earlier description of SYSTRAN, Hutchins and Somers (1992) described 5 types of dictionaries as part of the system. An “Idiom” dictionary with the translation for fixed expressions; a “Limited Semantics” dictionary which defines how the components of noun phrases are related; a “Homograph” dictionary to help in the disambiguation of homographs;

⁵ Term used by Senellart (2007) from the French “rattachement” meaning assignment.

an “Analytic” dictionary which includes syntactic information for words which diverge from the general syntactic behaviour; and a "Conditional" dictionary to ensure successful domain-specific lexical choices.

In addition to these dictionaries, general and domain-specific dictionaries are encoded. It is this type of dictionary which can be created by the user and facilitates the interaction with the in-built lexicon and rules, which are otherwise inaccessible to the user due to licensing agreements. These user dictionaries (UD) are fully customisable and overwrite existing in-built dictionaries. The user can specify not only the translation for a particular entry, but also its part-of-speech, how a plural is formed for irregular nouns or can give an entry a priority over different homographs.

One of the main criticisms of RBMT systems is the time required to develop new language pairs because rules and lexicons have to be devised and manually encoded. Recently, however, Surcin et al. (2007) have shown advances in the rapid development of new translation pairs at SYSTRAN. They argue that the source analysis and generation modules for each language can be reused when combining different languages. And as we mentioned earlier, source analysis accounts for 80% of the encoding and generation 10%. All that remains to be created, therefore, are the language-pair and direction-specific transfer rules. The crafting of grammatical rules requires some effort but the creation of lexical dictionaries is straightforward with the use of parallel corpora and semi-automatic extraction (Senellart, 2007). In addition, should new languages have to be integrated, all three modules incorporate language-family rules, more general than language-specific rules, which can be reused (Surcin et al. 2007; Hutchins and Somers, 1992).

Carl and Way (2003) argue that RBMT systems suffer from a "*knowledge-acquisition bottleneck*" (ibid: xviii) when trying to incorporate solutions for complex difficulties. Not only is there a need for highly specialised linguistic rules to facilitate disambiguation, often a new rule intended to improve a particular difficulty conflicts with other rules, resulting in considerable degradation of the quality. However, RBMT systems are incremental and deterministic in nature (Senellart, 2007). Users can easily validate and select improvements from different versions of the system and, due to the consistency of the results based on controlled resources, mistakes are easily pinpointed,

as opposed to data-driven architectures (see section 1.5.2.2 for further discussion of this).

It is worth mentioning that in June 2009, SYSTRAN announced the move towards a hybrid system, by incorporating statistical methods to their RBMT system (Laporte, 2009). The main advances described were a heavy use of statistics in the disambiguation rules during source analysis and the introduction of a Statistical Post-editing (SPE) module after the generation module (see 1.5.2.2 for a description of SPE).

1.3 -ING WORDS AND (MACHINE)TRANSLATABILITY

We describe here different translatability issues for -ing words. Let us start with an issue affecting clauses and phrases in general: structural ambiguity. This occurs when more than one interpretation of the syntactical relations between different clauses or phrases is possible (Quirk et al. 1985). For instance, in the example in Table 1.2, the -ing word *requesting* could be an adverbial phrase of manner modifying the main verb *close* (see possible re-writing 1) or a reduced relative clause complementing the direct object *dialog box* (see possible re-writing 2).

Structural ambiguity	Close the dialog box requesting a change.
Possible re-writing 1	Close the dialog box by requesting a change.
Possible re-writing 2	Close the dialog box that is requesting a change.

Table 1.2: Example of structural ambiguity

This type of ambiguity poses comprehension problems for humans, who might not be able to interpret the intended meaning correctly. Quirk et al. offer alternatives for avoiding such situations, e.g. changing the order of the clauses, supplying ellipted elements or using punctuation to mark the major clause boundaries, but it is usual to find them in real texts. If a structure is ambiguous for human readers, needless to say it will be so for an RBMT system. Thanks to the deterministic nature of these systems, however, we could expect that they consistently follow a default behaviour coded for such cases.

Nevertheless, often humans are able to interpret structural relations correctly by considering context. Clues can be found somewhere else in the text or logical relations can be applied to rule out impossible relations. However, MT systems cannot make use of world-knowledge and they can only focus on the clues provided within a sentence in

reference to grammar. For instance, in the example in Table 1.3, a human reader would rapidly know that the agent of the main clause is the user even if it is not present in the sentence. However, an RBMT system would have to resort to grammatical logic to disambiguate it. However, only semantic clues are provided in the sentence.

The password must be typed before accessing the account.

Table 1.3: Example of a passive sentence without an explicit agent and an implicit subject subordinate clause

The ambiguity for RBMT systems – and for humans – is maximised when the sentences violate grammatical rules. According to grammar (Quirk et al. 1985; Carter and McCarthy, 2006), we can infer the implicit subject of a subordinate clause by looking at the main subject, as they should be the same. However, this requirement is not always met and it is not unusual to find sentences like the following:

These credentials are stored in the credential database of the Information Server after configuring the UNIX target machines.
--

Table 1.4: Example of an ungrammatical use of implicit subjects

If we are to understand the sentence according to grammar, the subject who *stores the credentials* and the subject who *configures the UNIX target machines* should be the same. The first problem we find is that the agent of the passive clause is not made explicit and therefore we need to infer who the subject of the main verb is. Using world-knowledge, one could interpret this as a machine or program which stores the *credentials* and that it is a human who configures a machine. In this case, therefore, the subject of the main and subordinate clauses is not the same. Apart from confusing the reader as to who does what, we see that an RBMT system has no choice but to resort to a default assignment of subjects. The sentence does not only omit essential information for its correct interpretation, it also challenges the grammatical rule whereby subjects should be made explicit if they are not shared by both the main and subordinate clauses.

We present a more -ing word-specific issue here. The grammatical flexibility of -ing words results in words sharing the exact same form fulfilling different functions. For instance, the word *auditing*, in the examples below (see Table 1.5), can act as a noun (gerundial noun), as an adjective (participial adjective), or as part of the progressive tense or as the head of an adverbial phrase of purpose (gerund-participles).

TYPE	EXAMPLES
Gerundial noun	To perform auditing you must complete the following steps:
Gerund-participle	Server is auditing and logging.
	Steps for auditing SQL Server instances.
Participial adjective	When the job completes, BACKINT saves a copy of the Backup Exec restore logs for auditing purposes.

Table 1.5: Examples of different classes of -ing words

This characteristic presents difficulties mainly for RBMT systems, which need to differentiate between -ing word types for an appropriate analysis to be passed to the transfer modules. As mentioned in section 1.2.1, the analysis step relies heavily on part-of-speech analysis, where each word is assigned a tag for the morphosyntactic class to which it belongs. Tagging algorithms rely on the immediate environment of the word to be tagged in order to assign it the correct label (Jurafsky and Martin, 2009). If we examine the examples in Table 1.5, it is clear that the immediate environments for examples 3 and 4 are the same, that is, the word classes preceding and following the -ing word are the same for both sentences. But the -ing words belong to different classes.

Similarly, let us consider the following sentence:

The system backs up credentials and installing reports.

Table 1.6: Tagging ambiguity issue

If we only look at the syntactic structure, there are two possible interpretations of the sentence: (1) *reports* is a plural noun and the -ing word *installing* a participial adjective, or (2) *reports* is a verb in the third person singular and the -ing word a gerundial noun. Jurafsky and Martin (2000) list a number of ambiguous contexts for POS taggers and include plural nouns and third person singular verbs as one of the main examples. A full tree-parse of the sentence might help disambiguate the sentence. However, it might be the case that the parsing module relies on the POS tagger and therefore, this does not facilitate a correct analysis of the source. The RBMT system, therefore, is in a situation where it needs to make a choice, probably assigning a default tag for this situation.

Bernth and McCord (2000) maintain that words which can be assigned more than one POS tag are indicators of translatability issues. As a result, -ing words were penalised in their Translation Confidence Index for MT (TCI). Whereas this applies

generally, in certain cases, the frequency with which each POS appears varies and this can facilitate the disambiguation task. However, the TCI contains no mention of POS frequencies and how this could alleviate the translatability issue.

A TCI is a "*measure of the MT system's own confidence in its translation*" (Bernth, 1999b: 121). Thinking that MT users would benefit from the identification of potentially problematic structures before sending them for MT or after they had been machine translated, different researchers started working on TCIs in the 1990s (Gdaniec, 1994; Bernth, 1999b, Bernth and McCord, 2000; Bernth and Gdaniec, 2000, 2001; Underwood and Jongejan, 2001). In the same vein as the reported rule proportions for the different stages of SYSTRAN (see section 1.2.1) which devotes 80% of its coding to source analysis, all of the above authors emphasise the importance of source analysis as it is "*the most nondeterministic and most error-prone part of MT*" and "*errors made at this stage tend to carry over and influence later stages*" (Bernth and McCord, 2000: 90-92).

As mentioned, -ing words were included in the TCI lists. Underwood and Jongejan (2001) mention, somewhat tangentially, other structures where -ing words can be present, such as missing subjects and non-finite verbs. In Bernth and Gdaniec (2000) different uses of -ing words are specifically mentioned and coded for the English-German language pair, examples being adverbial clauses of manner directly introduced by an -ing word or gerundial nouns.

Based on the assumption that different RBMT systems are faced with similar problems, Bernth and Gdaniec produced a list which describes the grammatical features which can pose problems for MT systems: MTranslatability (2002). From the 26 rules proposed by these authors, 3 address the use of -ing words alone (ibid: 181-184).

Not only do (machine)translatability indicators include -ing words as a problematic grammatical feature for MT, authoring style guides and controlled languages (CL) also address them. In "The Global English Style Guide", Kohl (2008) dedicated a chapter to the use of -ing words while trying to raise awareness among technical writers by explaining the ambiguities and complexities behind this feature. 8 out of the most popular 10 CLs studied by O'Brien (2006), which aim at making source texts more

suitable for human comprehension and MT systems by eliminating complex and ambiguous structures, also included rules banning or minimising their use (see section 1.4.2 for a full description of CL).

Finally, we should mention a characteristic particularly important for generation modules. We have seen that English is a very flexible language when dealing with -ing words. However, other target languages might require a specific grammatical structure for the different cases, as a one-to-one correspondence for all the structures in which -ing words can occur is very rare. This means that RBMT systems need to differentiate between gerundial nouns, gerund-participles and participial adjectives, as well as the different structures within these classes.

To sum up, we could argue that -ing words can potentially pose structural ambiguities, reinforced by implication, POS flexibility and a lack of one-to-one correspondence between the SL and TLs for all the structures in which they can appear. The theoretical evidence and all the efforts to constrain the use of -ing words show that both scholars and practitioners perceive them as hindering RBMT performance considerably. However, very little empirical research has been done to quantify the problem posed by -ing words and few efforts have been made to date to test the methods that would best tackle the problem posed by this word category.

1.4 MACHINE TRANSLATION EVALUATION

Increased research on MT, including the emergence of new paradigms such as SMT, has made it possible to consider MT as a translation option for specialised domains or information gisting. However, the output of MT systems in general is still far from perfect. Evaluation, therefore, is of paramount importance in order to measure the quality and improvement but also to direct research towards the most urgent issues. Nowadays, we can find methods and algorithms for both human and automatic evaluation, both with benefits and drawbacks. Since our research on -ing words involves a substantial focus on evaluation, we will now review the relevant literature.

1.4.1 HUMAN EVALUATION

There are three main points to consider when performing a human evaluation of MT: who the evaluators are going to be, how they are going to perform the evaluation and what they are going to evaluate.

Humans understand their native languages and are fluent in them. They can decide whether a sentence sounds natural or not. In addition, end-users of translations are also humans. Therefore, it makes sense to use human judges to help in evaluating MT output. We observe two tendencies in the selection of evaluators in the literature: some studies use non-expert linguists whereas other studies use experts. The choice is usually influenced by pragmatism rather than methodological rigour. Experts very rarely volunteer to participate in evaluation tasks as they can take a long time and may require iterative cycles, and therefore, evaluators must usually be compensated for their involvement. The evaluation, then, becomes costly, especially if a large number of evaluators is required (see section 2.2.1.3 in Chapter 2 for further discussion). Availing of non-experts free of charge is easier. They are usually university students (Shubert et al. 1995; Babych et al. 2009) or random people who access the evaluations online (Callison-Burch et al. 2006). Recently, Zaidan and Callison-Burch (2009) explored the possibility of obtaining evaluations from paid evaluators by posting the task in a virtual marketplace (Amazon's Mechanical Turk). By presenting the task as a job, the reliability of the answers might be higher. 3,873 units were evaluated by 115 evaluators in 4 days. Overall, they calculate having to pay an hourly rate of \$1.95, much lower than the rate of a professional evaluator. The authors report that identifying and blocking evaluators who perform poorly is easy if they submit a relatively large number of evaluations. Random or inconsistent answers can be noted by inspecting a number of responses. However, no requirements are reported to have been established when recruiting evaluators, which questions their capacity to perform the task.

The use of non-experts carries risks. If we use experts, i.e. people who have completed training in languages or translation, it could be argued that they possess certain language, textual, subject, cultural and transfer competences (Neubert 2000: 5-10). This puts them in a good position to provide a reliable assessment on the different aspects of language, such as grammaticality, fluency and accuracy. This cannot be

ensured when availing of non-experts. In the case of experts, however, a budget must be allocated.

Machine translation quality evaluation methodologies and the attributes required of human judges have been discussed since the 1970s. After the acquisition of a version of SYSTRAN and in order to embark on the EUROTRA project (1978) to develop its own MT system, the European Commission needed recommendations for evaluation. The “Van Slype report” (Critical Methods for Evaluating the Quality of Machine Translation 1979) was commissioned. It aimed at providing an overview of MT evaluation and at advising the EC on evaluation methodology and application. The report established a framework which, among others, detailed definitions of translation quality attributes to be judged by human evaluators, e.g. comprehensibility, fluency, accuracy. It was not publicly accessible until 2003 (King et al. 2003).

Between 1993-1999, the EC set up EAGLES (Expert Advisory Group on Language Engineering Standards) to propose standards, guidelines and recommendations for good practice in the evaluation of language engineering products. 1996 saw the first guidelines published.⁶ In April 1999 the EAGLES Evaluation Working Group published the EAGLES 7-step recipe for evaluation where steps for the definition of the evaluation purpose and intended users, quality characteristics to be measured, metrics to be applied and execution design were recommended.⁷ Both documents established the backbone of evaluation methodology. However, neither further investigated the quality attribute definitions. The standard does not establish how to measure the quality of the output. Although there are mentions of *usefulness* or *adequacy*, for instance, no definition is given for the terms.

Despite the efforts to establish a well-defined methodology and attributes, no standard was agreed upon and studies tended to redefine the existing ones to better suit their context. Through collaborative work in the ISLE project (International Standards for Language Engineering) and with funding from the European Union, the National Science Foundation in the USA and the Federal Office for Education and Science

⁶ For EAGLES Guidelines see: <http://www.ilc.cnr.it/EAGLES96/browse.html> [Last accessed on 31.10/09].

⁷ For EAGLES 7-step recipe see: <http://www.issco.unige.ch/projects/eagles/ewg99/7steps.html> [Last accessed on 31.10.09].

(OFES) in Switzerland, FEMTI (Framework for the Evaluation of Machine Translation in ISLE) was created (King et al. 2003).⁸ FEMTI emerged as an attempt to “gather into one place the accumulated experience of MT evaluation, and to describe it in such a way that future evaluators can consult and re-use this experience easily” (King et al. 2003:1). The framework is easily accessible. Given the purpose of the evaluation, it suggests the best suited attributes, together with their definitions from different authors, and the methods to use, referring the user to past case-studies, e.g. scales, performance tests, comprehensibility tests.

FEMTI offered a recompilation of attributes and tests to perform a targeted evaluation. However, from the beginning of the discussions, two main parameters, which have evolved over time, seemed to emerge. Pierce stated that “the two major characteristics of a translation are (a) its intelligibility, and (b) its fidelity to the sense of the original text.” (1966: 67). Intelligibility is a target-text quality, whereas fidelity accounts for the equivalence of the source and target texts in terms of informativeness. The final report of the ARPA MT evaluation methodology (White et al. 1994) also proposed an attribute to account for the information shared by the source and target texts which they called *adequacy* and an attribute to evaluate the quality of the target text, *fluency*.⁹ We note that even with different names, two aspects are recurrently considered to evaluate machine translations. We will return to the definitions behind these attributes, and in particular that of fluency, in Chapter 2 section Evaluation Attributes.

Recent large-scale machine translation evaluation campaigns continue to use the same parameters and seem to have reached consensus in the use of two particular terms, “fluency” and “accuracy” (LDC, 2003; Callison-Burch et al. 2007), where evaluators are asked to assign sentences a value on a 5-point scale. However, disagreement seems to be emerging again, as this is considered a time-consuming task (Callison-Burch et al. 2008), and ranking of sentences from different systems is starting to gain popularity. Despite the variation of evaluation attributes, a standard evaluation unit for human evaluation is observed: the sentence. A sentence provides the evaluator with a

⁸ For FEMTI online see <http://www.issco.unige.ch:8080/cocoon/femti/st-home.html> [Last accessed on 31.10.09].

⁹ They suggested including a comprehension test.

manageable unit with complete meaning - it has a subject and a predicate. It is debatable, nevertheless, whether the evaluation of isolated sentences is an adequate unit to assess any attribute measuring understanding or fidelity as textual coherence and cohesion, as well as extra-textual elements, greatly help human readers construct the meaning of a text. Also, whereas the evaluation of sentence units provides a good overall review of the systems, it falls short of providing researchers with information as to where to concentrate their efforts. From a development point of view, and particularly working with RBMT systems, researchers would greatly benefit if evaluators could provide more granular information as to what structure or word requires improvement. A methodology or even a typology for such an evaluation has not been examined to date.

Human evaluation has been accused of being subjective, inconsistent, time-consuming and expensive (Coughlin, 2003; Popescu-Belis, 2003; Callison-Burch et al. 2006; Hamon et al. 2007). Coughlin adds that "*[r]ating translation quality is both tedious and repetitive*" (ibid: 2). Therefore, she suggests, due to the different working tolerance of individual evaluators towards these conditions, the results of the evaluation are compromised. However, there are no alternatives to performing human evaluations if we aim at obtaining a degree of informativeness on quality and the identification of incorrectly translated structures.

1.4.2 AUTOMATIC EVALUATION

Machine translation output has traditionally been evaluated by humans because they have the linguistic competence and can deal with abstract concepts such as quality. We established in section 1.4.1, however, that performing a human evaluation is time-consuming. It is, in addition, expensive, as, where expert linguists or translators are required, they must be compensated for their time. This is an obstacle for the MT development community, as they require quick and continuous evaluations to measure the effect of the modifications they continuously introduce to systems. Similarly, it is troublesome for monitoring ongoing production cycles where the need exists to measure the effects of the changes implemented along the information production workflow.

With the aim of automating this task, researchers proposed the use of automatic metrics which are able to distinguish between a good and a bad translation. Most of these metrics are based on the premise that "*[t]he closer a machine translation is to a professional human translation, the better it is*" (Papineni et al. 2002a: 311) with the result that they compare MT output against reference translations created by humans. One could argue that even though such metrics are termed automatic, they are, in reality, semi-automatic, in the sense that one or more human-generated references must still be acquired for the translations to be evaluated (see Chapter 2 for discussion of reference translations).

This is true for the widely used automatic metrics such as BLEU (Papineni et al. 2002a) or GTM (Turian et al. 2003), which are string-based, that is, they compare strings of surface forms of the machine translated text against the human reference text. Recently, attempts have been made to go beyond the use of reference strings. For example, researchers are developing metrics which extract syntactic information from the source text and the target text and measure the difference (Liu and Gildea, 2005; Owczarzak et al. 2007a, 2007b). The information extracted can go as far as complete parsed-trees with dependency relations and morphological analysis of each word, e.g. person, tense and aspect of a verb.

Automatic metrics greatly facilitate system comparisons, at a text or sentence level. Taken in isolation, however, the scores, which should reflect human judgement, have no meaning and give no information about the actual quality of the machine translation. The meaning of an increase of a BLEU point in terms of quality has not been described yet. Nonetheless, often increases of points are reported as achievements (Och, 2003; Callison-Burch et al. 2006; Huang and Papineni, 2007; Niehues et al. 2009). One could assume that with the advance in more grammatically-driven automatic metrics we will be better able to assess the meaning of results.

Many studies showed that automatic metrics correlated with human judgements (Doddington, 2002; Coughlin, 2003) and the research community was quick in adopting these fast and cheap evaluation tools. However, other studies have emerged that challenge this assumption. For instance, some claim that the metrics correlate well at text level, but not at sentence level (Blatz et al. 2003; Kuleska and Shieber, 2004) or prove that they do not account for syntactic variability (Och et al. 2004).

Callison-Burch et al. (2006) identified the reasons why BLEU may not always correlate with human judgements. They claimed that BLEU allows for excessive n-gram variation for sentences with the same score. As long as the number of each n-gram level is the same, the BLEU score does not vary, regardless of the position in which the n-grams appear and the lexical items they contain. Moreover, the metric cannot account for any variation that is not present in the reference translations and does not apply any weight difference for content words and function words. This results in the MT output and two variations in Table 1.7 scoring equally, although clearly divergent on quality. In addition, systems with different translation strategies, data-driven vs. rule-based, for instance, seem to receive different scores, indicating that they should not be compared directly through these metrics (Callison-Burch et al. 2006; Koehn and Monz, 2006).

	Example
Reference 1	Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.
Reference 2	Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.
Reference 3	Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.
Reference 4	Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.
MT output	Appeared calm when he was taken to the American plane , which will to Miami , Florida .
Variation 1	Which will he was , when taken Appeared calm to the American plane to Miami , Florida .
Variation 2	Appeared calm black he was helicopters to the American plane , which will to Miami , Florida .

Table 1.7: Set of four reference translations, an MT output and two variations with the same BLEU score from the 2005 NIST MT Evaluation (from Callison-Burch et al. 2006)

We have seen that MT evaluation is not a standardised process. Depending on the purpose and available resources, one can choose between a human evaluation or an automatic evaluation. Both options come with additional complexities. Human evaluations have restrictions on the amount of data that can be judged, call for well-defined attributes and must cope with subjectivity. If methodological rigour is applied, however, we argue that it is possible to design a targeted task with highly informative results. Currently used (string-based) automatic options return fast and robust results and are useful for comparison purposes. The downside is that a human qualitative assessment is then necessary if we are to gain any insight into the quality

and quantity of the improvements and degradations. We will return to this concept in Chapter 2.

1.5 MACHINE TRANSLATION IMPROVEMENT

MT evaluation detects weaknesses in MT systems, setting the path for further research and development. In RBMT systems, improvements can be made by modifying the system architecture, and specifically, the so-called core rules. However, it could be argued that it would be daunting to confront over 30 years of code (in the case of Systran) and try to understand all rule combinations and interactions. Yet, not all users have access to these rules due to licensing agreements, which leads to black-box scenarios like the one examined in this study. In such cases improvements can be pursued either before the text is submitted for translation or after it is machine translated. We call the former the pre-processing stage and the latter the post-processing stage.

1.5.1 PRE-PROCESSING

Pre-processing encompasses any method used to manipulate the source text to suit the MT system better. Techniques range from simplifying strings, or avoiding particular structures a specific system cannot handle correctly, to introducing additional information to help a correct analysis. The best known approach used during pre-processing is Controlled Language. In addition, tagging and source re-generation are emerging as new pre-processing methods.

1.5.1.1 CONTROLLED LANGUAGE

Hauser et al. (2002: 1569) describe the human faculty of language as follows:

human faculty of language appears to be organized like the genetic code – hierarchical, generative, recursive, and virtually limitless with respect to its scope of expression

Natural languages offer limitless possibilities of expression, which often leads to complex and ambiguous structures, as we have seen with the -ing form in English in section 1.3. Therefore, Huijsen (1998) claimed that understanding natural language might pose some difficulty for both humans and computers. Humans use language to communicate, sometimes directly and other times via translation. Humans can find

translation difficult; as Arnold et al. (1994) described, not only a good command of the source and target languages and their corresponding meaning is required but also a good insight into culture, customs, and expectations, as well as certain expertise on the subject matter of the document to be translated. As for computers, Arnold (2003) claimed, translation is difficult because (1) as mentioned before, translation is inherently difficult and (2) computers have particular limitations for learning, reasoning, and dealing with high numbers of combinatorial possibilities. This is because they do not have access to real world-knowledge. Moreover, the human-machine interaction is done through programming languages, which act as intermediaries between natural language and binary language. All these limitations greatly affect the capacity to process text and translate, at least if this is to be done by modelling the human process. In an attempt to overcome language difficulties and enable effective communication, scholars proposed the use of Controlled Language (CL), which is defined as:

an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style (Huijsen, 1998 :2)

AIM OF CL

The ultimate objective of CL is to limit the ambiguities of language. Basically, CL aims to promote communication, and therefore, its application ranges from learning a language to processing text using computers. In fact, the first CL was developed as a standard for learning English. In the 1930s, Charles Kay Ogden developed Basic English, a set of 850 words and several rules to restrict inflection and derivation. By using these limited resources, a non-native English speaker was supposed to be able to communicate in this language (Ogden, 1930).

CL was not implemented for the translation process until the 1970s, when the Caterpillar Corporation developed Caterpillar Fundamental English (CFE) (Kamprath et al. 1998). Their example was followed and different CLs were developed, such as Smart's Plain English Program (PEP), the International Language for Servicing and Maintenance (ILSAM), Clear and Simple English (CASE) and Perkins Approved Clear English (PACE) (Nyberg et al. 2003). PEP set the foundations for the CL which would be used at Clark, Rockwell International and Hyster (ibid). ILSAM inspired AECMA SE, IBM, Rank Xerox, and Ericsson Telecommunications (ibid). Gradually,

large companies in the aeronautical industry (Spaggiari et al. 2003), in the automotive industry (Pym, 1988; Almqvist and Sâgvall Hein, 1996, 2000), in IT companies (Bernth, 1998; Wells Akis and Sisson, 2002; Roturier, 2004), in the telecommunications industry (Adriaens and Schreurs, 1992) and even in financial services (Dervišević and Steensland, 2005) have been adopting CL.

In theory, CLs offer significant advantages for companies. Although, as we will see in the following sections, the benefits have not always been demonstrated, here are, according to Muegge (2007), the five players who could benefit from the implementation of a CL in a localisation environment:

1. Users of software and documentation. Improved readability for this group means an increased usability of the product as well as a reduction in the amount of support-related costs.
2. Authors of technical documentation. The use of a CL creates a framework which structures the work of technical writers making it faster and even able to provide quality metrics.
3. Translation technologies such as translation memory (TM) tools. More consistent texts translate into an increased rate of absolute and fuzzy matches which means lower translation costs.
4. MT systems. Because source texts would become less complex and terminology would be controlled, the quality of MT-generated translations would improve. As a consequence, more text could be translated using this resource.
5. Translators.¹⁰ From an increased quality in MT output, post-editing would become faster, with quicker turnaround times.

Using CL, texts received for translation would be more comprehensible and unambiguous, and would contain a controlled terminology. As Hurst (2006) pointed out, clarifying an ambiguous source sentence for a translator can be time-consuming.

¹⁰ Muegge refers to the professionals responsible for editing MT output as translators. However, note that in this thesis these professionals are referred to as post-editors, whereas the professionals who translate from the source text, using translation memories when applicable, are referred to as translators.

In order to check whether the translator's interpretation is correct, the question needs to go through a chain of project managers to the original writer. Once the answer is known, it must go back to all the translators, to avoid any misinterpretation in all target languages, as well as to the source language editors, to modify the source texts.

We can distinguish two main foci for CLs in a localisation environment. First, the improvement of readability and comprehensibility for humans and, secondly, the increase of text processing possibilities for computers, also called (machine)translatability. Based on these objectives, CLs have been divided into two categories: human-oriented CLs and machine-oriented CLs (Huijsen, 1998; Nyberg et al. 2003).

There is debate regarding the compatibility of these two goals. Bernth and Gdaniec (2001) showed that after applying MT-oriented CL rules their text corpus improved in clarity and translatability but reduced readability. However, Reuther (2003) concluded that the rule sets required for both readability and translatability are very similar and that, in fact, "*translatability ensures readability*" (ibid: 131). For readability, rules dealing with lexicon and ambiguity proved the most important, whereas rules dealing with ellipsis and typography had little impact. In contrast, translatability relied mostly on rules dealing with ambiguity and ellipsis, whereas it did not depend on lexical rules. A recent study, mainly focusing on rules addressing readability but also including MT-oriented rules suggested that controlled texts were "*easier to read, are viewed more favourably, and encourage better retention of keywords*" (Cadwell, 2008: 50). A follow-up study, on the contrary, limited these results by showing that CL rules might be beneficial in terms of readability and acceptability for complex texts but not for easy texts (O'Brien, forthcoming). More comprehensive studies on the relation between readability and translatability are required to shed light on these contradictory findings.

One must also bear in mind, however, that the concepts of readability and translatability themselves are not agreed upon. As Cadwell (2008) mentions, authors who try to measure the improvement CLs bring to readability and/or translatability do not use a standard definition of those terms. What is more confusing, each seems to either coin new terms with study-based definitions or re-use existing concepts with a different understanding of them. The confusion is such that authors contradict one

another. For instance, Reuther (2003) claims that readability is a subset of translatability, whereas Hargis (2000) understands translatability as just one level of readability.

CL AND TRANSLATION

Nowadays technical translation is often carried out either by using translation tools such as translation memory systems (TM) or by machine translation (MT). As for TM tools, the more consistent the source text, the more efficient the translator becomes. Nyberg et al. (2003) claim that since CLs reduce variation, they can increase reusability of source texts quite significantly, thus lowering overall translation costs. Moreover, thanks to the additional possibility of addressing MT weaknesses, they also claim that CLs are very good candidates for improving MT output (ibid). Aware of this potential of CLs, many companies such as IBM, Diebold or Symantec have introduced them into their document production workflow (Bernth, 1998; Moore, 2000; Roturier, 2004).

One of the most beneficial deployments of CL for MT is that developed at Scania where the analysis module of the MT system is shared with the CL checker (Sågvall-Hein, 1997). By doing so, the technical writers can focus on improving the specific sentences or chunks that pose problems for the system. However, the use of a CL to try to enhance MT output when using a commercial, proprietary MT system can be impeded due to the “black-box” effect. Once the phenomena which increase translation quality the most are identified, sentence length or passive voice to mention but two, more precise knowledge of the MT system is required to address its particular weaknesses. Moreover, one should remember that the company providing the MT systems will release new versions of it and changes will impact on the work done to address the identified weaknesses.

But to what extent can the implementation of CL improve MT output? Several studies have been carried out to try to measure the actual increase in translatability when CLs are applied. Spyridakis et al. (1997) asked Spanish and Chinese translators to translate procedural documents written either in natural English or in Simplified English (SE).¹¹ They concluded that SE appeared to improve the translation in certain

¹¹ ACEMA SE was officially renamed ASD Simplified Technical English in January 2005.

cases, particularly for Spanish. This could mean that SE is more efficient when dealing with linguistically similar languages (ibid). De Preux (2005) also carried out a quantitative analysis. She focused on counting error-severity scores. The results suggested that although the number of errors did not decrease with the implementation of CL, their severity was reduced (ibid). In a different study, a significant improvement on the output of a commercial MT system using CL was reported (Roturier, 2004). Output was classified as excellent, good, medium and poor. Excellent output is defined as being ready for review.¹² In Roturier's context, poor output is discarded and the source sent to translators for traditional translation. Good and medium quality output is sent to post-editors. Overall, excellent output doubled for all languages and medium quality examples decreased considerably. It should be remembered that, as Ryan (1988), backed by O'Brien (2006), claims, the efficiency of MT should be measured taking into account the time and effort required for the whole translation process, which includes post-editing, unless the translation is for information gisting.

CL AND POST-EDITING

As Allen (2003) reports, CLs and post-editing (PE) are often combined to enhance translatability and therefore improve MT output quality, thus making PE faster (see section 1.5.2 for a full review of PE). In the same line, Krings (2001) claims that medium quality output might take longer to repair than to translate from scratch and, therefore, CLs should focus on eliminating medium quality output. O'Brien (2006) devoted her PhD thesis to trying to draw relations between CL and PE effort. She reported that the occurrence of negative translatability indicators (NTIs), that is, specific translatability issues for MT, does not necessarily always increase PE effort. Her study supports the assumption that controlling the input reduces PE effort. However, she points out that different types of NTIs have a different impact on PE. Of particular interest for our study, she found that "gerunds" belonged to the group of NTIs which appeared to cause higher post-editing effort, although the definition of gerund used is only a subset of our broader definition of -ing words.

Allen and Hogan (2000) see PE and CL authoring in the same light, arguing that CL authors and post-editors perform similar changes to the texts. It could be argued that

¹² The review is the last stage in a translation cycle where a senior linguist signs off on quality and accuracy of a translation before returning it to the client.

the “mirrored grammatical structures” found in the controlled source text and MT output (MT systems tend to follow the source structure closely) create a controlled target (Way and Gough, 2005). However, that both CL authors and post-editors perform similar tasks seems debatable. We should at least note that the type of text they work with is different: whereas authors or editors deal with natural and mostly grammatical texts, post-editors repair artificial and usually ungrammatical texts.

PORTABILITY OF CL

Given that the goal of CLs for large companies is mainly readability, comprehensibility and/or translatability, one might assume that rules are portable across companies. However, after analysing eight different CL rule sets for English, O’Brien (2003) found that the rules varied significantly. The rules chosen for each CL seemed to depend not only on the purpose of the CL (readability and/or translatability), but also on the weaknesses of a particular MT system, the specific translation difficulties between a language pair, corporate writing guidelines and the developers’ concept of CL. A company wishing to implement a CL, therefore, would need to invest in the analysis of their source texts, target text requirements and, if appropriate, MT system, which constitutes a significant endeavour. In order to set a starting point, O’Brien gathered a list of what she considered “*the most important rules for improving machine translatability*” (ibid: 111), classified into lexical, syntactic, semantic, text structure and pragmatic rules. In the same line, Bernth (1997, 1999a), Lehtola et al. (1998) and Nyberg et al. (2003) emphasise the importance of tailoring a CL to the MT system in use in order to improve its translation.

Shared efforts have also been made to attempt to find CL rules that are useful across companies. O’Brien and Roturier (2007) for instance, presented a comparison of the implementation of a CL for the German/English language pair in an IT context. The experiments showed shared benefits from controlling misspelling, misuse of the question mark, semi-colon and double hyphen, long sentences and personal pronouns with no antecedents. More generally, and as a starting point for those wishing to implement a CL, the authors recommended starting by focusing on those rules addressing common source analysis problems, as these are the ones that seem to have a shared benefit across experiments. Rochford (2005) studied whether a CL developed with a particular MT system in mind could be reused and the same effects obtained

when applied to a different MT system. She reported that whereas the best results were obtained when the targeted MT system was used, benefits could be seen for the other two systems. It should be noted, however, that the author did not provide any information on the characteristics of the MT systems tested, due to confidentiality restrictions. Her work builds on the literature that claims that MT systems share common weaknesses and, therefore, CL rules defined with MT in mind benefit different systems, albeit to different degrees.

From O'Brien's findings, it could be concluded that the implementation of a CL is highly dependent on the particular setting of a company. Existing writing guidelines, the translation workflow (human and/or machine translation, computer-assisted translation (CAT) tools) and expected final quality of texts will influence the decisions about the rules to be included in the CL and the acceptable strictness. Within the machine-oriented CLs, Huijsen (1998) and Nyberg et al. (2003) presented a further division: loosely-defined CLs and strictly-defined CLs. The former intend to improve the source text and the authoring rules are therefore not very precise. Perkins Approved Clear English is an example: *"keep sentences short"*, *"order the parts of the sentence logically"* and *"avoid strings of nouns"* are three of the "Ten rules" used by this CL (Pym, 1988: 88-89). On the other hand, strictly defined CLs aim at producing the most suitable source text for a particular MT and, therefore, try to be as specific as possible. Caterpillar's CTE in combination with the KANT system is an example. Rules such as *"repeat the preposition in conjoined constructions where appropriate"* and *"Both clauses in complex sentences using subordinate conjunctions must contain a subject and a verb; if the subordinate conjunction is removed, the subordinate clause should be able to stand alone as a simple sentence"* are part of this CL, and are clearly more specific (Nyberg et al. 2003).

From the point of view of technical writers, loosely defined CLs grant greater writing freedom than strictly defined ones. These CLs comprise a more succinct list of rules which tend to be more general, and as a result, easier to remember. Nevertheless, the resulting texts incorporate the rules in a broad sense, as they are often open to the author's interpretation and the potential benefits might not be self-evident. For instance, the rule *keep sentences short* might be interpreted differently by each author. One might think it refers to paragraph-length sentences, whereas another might put

additional effort to shorten sentences to as few words as possible. A common CL rule limits the sentence length to around 25 words.

On the other hand, strict CLs leave little room for creativity and can constrain significantly the allowed vocabulary and grammatical structures. Due to the granularity of the rules, the list would tend to be longer and more difficult to remember. For instance, the rule “*avoid gerunds used as nouns*” might require some effort from authors, as -ing words appear in different contexts. The problem is exacerbated by the likelihood that a technical writer is not a linguist by training, but rather has stepped sideways in his/her career as an engineer, for example, into authoring. With no formal training in grammar, writers find it difficult to apply rules expressed in strict grammatical terms. The level of strictness can even go as far as to present the user with the possible POS and/or terms that can follow a word (Schwitter et al. 2003). This strictness, which compromises writing freedom, could result in completely processable texts for which an MT system would generate a perfect output, but which might be unacceptable to technical writers.

CL AND AUTHORING

The people responsible for the successful implementation of CLs are technical writers, and therefore, it is essential to obtain their consent for a successful deployment. Nyberg et al. (2003) claim that writers might argue that it is difficult to remember and to conform to the rules. Moreover, complaints may also arise from the restriction on creativity experienced when using CLs (ibid 2003). Several studies have confirmed this opposition and suggested involvement of writers in the development stage and training. Roturier (2004: 12) emphasises the need to make editors understand the issues faced during the translation process and the impact of the rules as this might encourage them to comply with the CL. Wells Akis and Sisson (2002) also list the need to involve the writers and editors in the customisation of CL as a key factor. However, Moore (2000) reports resistance to CL from authors despite their participation in the development. The advantages of using CLs should be pointed out: document consistency and reduced ambiguity, which lead to increased translatability and a faster, and potentially cheaper translation process (Wells Akis and Sisson 2002). It should also be acknowledged, though, that the introduction of a CL implies an additional stage in the authoring process which could also imply the need for additional time (ibid).

Given the difficulty in conforming to CL rules, different tools have been developed as an aid for authors. Writing tools such as the writing interface and the look-ahead editor (ECOLE) presented by Vertan and v. Han (2003) and Schwitter et al. (2003) respectively are an example. This type of editor guides writers through the production of a sentence offering word or category options as they write. The programs consist of strictly defined sentence structures and lexicons. ECOLE also generates a paraphrase that explains how the system interprets the current input. The resulting text is tightly controlled, and therefore, the quality of the translation is high. Vertan and v. Han, and Schwitter et al. however, do not mention whether the time a writer needs to complete the job increases, and if so, by how much.

CL checkers are another widely used tool. Nyberg et al. define these as “*a specialized piece of software which aids an author in determining whether a text conforms to a particular CL*” (Nyberg et al. 2003:251). These authors classify CL checkers as proscriptive and prescriptive. Although both flag unacceptable structures or terms, the former include rules on unacceptable structures and the latter on all possible correct structures.

Whereas most CL checkers were developed to comply with a particular CL within companies, nowadays more general checkers, such as acrolinx’s acrolinx IQTM, across’ crossAuthor Linguistic, SMART Communications’ MaxIT STE or Tedopres’ HyperSTE are available which offer the possibility of customising rules.¹³ To mention an example of CL checkers, acrolinx IQTM will be briefly presented. acrolinx IQTM is an “*Integrated Quality Assurance Tool*” (acrolinx) for technical documentation provided by acrolinx. It guarantees conformance with corporate writing guidelines by a full customisation of the product. acrolinx IQTM works by combining knowledge of acceptable and unacceptable structures. It could be argued that it is both prescriptive and proscriptive. For instance, it can be asked to flag all instances where -ing words appear (because in general they are considered problematic) except where an -ing word

¹³ acrolinx IQ (with natural language processing technologies developed at the German Research Center for Artificial Intelligence - DFKI):
http://www.acrolinx.com/iq_overview_en.html
 crossAuthor Linguistic (based on IAI’s CLAT):
ftp://ftp.across.net/fact_sheets/fact_sheet_iai_en.pdf
 HyperSTE: <http://www.tedopres.com/en/products-services/simplified-technical-english/>
 MaxIT STE: <http://www.smartny.com/simplifiedenglish.htm>

is followed by a determiner. Despite some opposition from writers, checkers have proven very efficient in improving readability, consistency and translatability (Roturier, 2004; Wells Akis and Sisson, 2002; Pym, 1988).

It is worth mentioning here that so far CL checkers have focused on the sentence level only. In 2006, Bernth presented a report on the move of IBM's EasyEnglishAnalyzer to a discourse level, but no further accounts have been given of this development since then. It uses document structure tags to establish what section of the document the text is in and checks it accordingly. Text size and organisation, and correspondence of paragraphs and topics, for instance, are issues which were reportedly being addressed.

Despite the help checkers provide, it should be remembered that the fine-tuning and maintenance of a checker can be expensive. Moore (2000) reported such a case. She presented 25% savings on a translation budget of \$100,000 by implementing CL at Diebold. However, a higher saving rate was required to maintain the checker. Further development of the software was consequently abandoned.

Finally, it is worth mentioning a tool suggested by Allen (1999): the authoring memory (AM). This idea takes advantage of the strong points of both CLs and TMs. CLs restrict the possibilities of usage and homogenise structures and lexicon. TMs are based on repetition. The more consistent and repetitive a text, the more likely matches will appear. Allen (ibid), therefore, proposed a tool to store source texts written to conform to a CL to enhance reusability. As an experiment, he re-authored 4 operation and maintenance manuals written using Caterpillar Technical English imitating the concept of TM. He reported that by the fourth manual only 25% of the text needed to be authored. No further development of the idea has been reported since. This could be due to the advances in content management systems (CMS), which ensure re-usability in real-life workflows.

CL AND USERS

CLs help authors develop texts that are easy to understand and translate, but their effectiveness in achieving this goal has not been questioned very often. When we compare the pre- and post-CL examples in Table 1.8, we can see that CL is an

effective tool for increasing clarity. However, what difference does a CL make to the user's daily experience?

Natural language	Controlled language
<p>Remove screws holding the blower and pull the blower from the cabinet.</p> <p>Before the screws are installed to the blower, a new blower is pushed back into the cabinet.</p>	<p>1 Remove screws from the blower.</p> <p>2 Pull the blower from the cabinet.</p> <p>3 Push a new blower into the cabinet.</p> <p>4 Secure the blower with screws.</p>

Table 1.8: Example of natural and controlled language sentences (taken from Quah 2006: 48)

Very few studies have addressed readability and comprehensibility (Shubert et al. 1995) and only one has focused on user satisfaction (Roturier, 2006). In the experiment carried out by Shubert et al. a group of native and non-native speakers of English were randomly distributed one of two procedures (one more complex than the other) written in natural language or using a CL (Simplified English). They were then asked to answer a set of questions about the content and to specify where in the text this information was found. The results suggested that the more complex procedure improved in comprehensibility and retrievability when written according to CL rules whereas the simpler procedure did not. Roturier focused on user satisfaction when presented with machine translated natural language documents (MT) and controlled language machine translated documents (CL + MT). The users who had visited one of the online technical support documents processed for the experiment were invited to answer a set of five yes/no questions. These provided information about comprehensibility, usefulness and acceptability. It was found that the *"hypotheses [on improved comprehensibility and usefulness] formulated based on the expert opinions of translators had to be rejected"* (Roturier 2004: 195). This was attributed to the possibility that end users have different expectations and reasons for reading texts compared to translators (ibid). It should be pointed out that this study and other user satisfaction studies (Lassen 2003; Jaeger 2004) report very low response rates, finding it very difficult to recruit people for the studies.

It could be concluded that the improvement in user satisfaction in the experiments analysed depended on the complexity of the source text and the TL. The use of CLs proved not to hinder comprehensibility, but failed to increase satisfaction in many cases. This finding disagrees with the general expectations of a CL. So far no comprehensive study has been carried out to show the impact of CLs on users where significant response was obtained.

CURRENT APPROACHES TO -ING WORDS IN CLs

Seeing that they were listed in translatability reports, -ing words are often allocated a rule in CLs. As mentioned, O'Brien (2003) found that six out of the eight CLs she analysed shared a rule which recommended avoiding gerunds. PACE proposes avoiding all -ing words (Nyberg et al. 2003). The customised rule at Symantec at the outset of this research behaved similarly, suggesting to writers not to use -ing words, even if some structures were accepted by the checker (Roturier, 2006). Dervišević and Steensland (2005) write that AECMA Simplified Technical English (AECMA 2004) does not allow the use of either gerunds or present participles, with the exception of certain technical terms.¹⁴ Following the trend, these authors also ban gerunds and present participles in the proposed CL for IFS (ibid).¹⁵ The Microsoft Manual of Style for Technical Publications (MSTP) is less restrictive (Microsoft Corporation 1998). Instead of banning the use of gerunds, it cautions that they can be ambiguous and be problematic for translators, leaving the choice of whether or not to use them to the writers.

Despite the emphasis on banning gerunds and present participles in general, it might be the case that not all structures where -ing words occur are equally problematic. In his Global English Style Guide, Kohl (2008) set to describe which -ing words pose problems and proposed to address only these. This author devoted a chapter to the use

¹⁴ Dervišević and Steensland (2005) define gerunds as “the ing-form of a verb, when functioning as a noun” (p120) and present participles as “the ing-form of a verb, when functioning as an adjective” (p121). The former coincide with what we classify as gerundial nouns and -ing words used at the beginning of titles, whereas the latter refer to pre-modifiers in our classification. They claim that by banning these two categories “no ing forms are allowed if not in the IFS Term Database” (p120). However, there is no mention of -ing words introducing adverbial clauses, post-modifiers, progressive tenses or other -ing words with a referential function, all of which are included in our classification.

¹⁵ IFS is “one of the world’s leading providers of component-based business software developed using open standards”. <http://www.ifsworld.com/>

of -ing words. He argued that -ing words were problematic for three reasons: (1) because they fulfil different grammatical functions which might not have an equivalent construct in the target languages, therefore confusing non-native speakers of English; (2) because native speakers use -ing words ungrammatically or do not punctuate -ing words correctly in certain contexts; and (3) because -ing words can be ambiguous in some contexts.

SUMMARY OF CL

From this section on CL, we can conclude that the purpose and degree of constraint is highly dependent on the particular working scenario and expectations of the group implementing the CL. We propose that CLs should help authors write consistent, unambiguous and grammatical English. By meeting these three requirements, we expect that source text readers and translators would understand texts better, and the text would be more translatable. Consistency would also allow for a more efficient use of TMs. Additionally, MT systems would also benefit from meeting the three requirements mentioned above. RBMT systems require that the source text conforms to the source language grammar, and therefore, the analysis is strongly hindered by grammatical violations. Secondly, the less ambiguous the source text is, the easier the MT analysis step would be. Finally, a consistent input would make it possible to customise the system (through user dictionaries, specific settings) for a better output quality.

Although the approach to CL depends highly on the particular needs of each company/workflow, a number of considerations must be made when using CL to compensate for the weaknesses of a particular MT system. An MT-dependent CL might lead to a high number of very strict rules which might be hard for technical writers to familiarise themselves with. The amount of training and skills necessary to master the CL might increase with the strictness of the rules. Moreover, the portability of the rules must be contemplated. The rules could become obsolete in newer versions of the MT system or not be transferable should the company decide to change the MT system. Work can be done to make the source text more MT-friendly, but CL rules might not always be the best mechanism for this.

OTHER METHODS

To the best knowledge of the author, no other attempts at controlling -ing words have been made. However, some efforts to transform the source text before machine translation must be acknowledged.

Babych et al. (2009) explored the possibility of creating automatic re-writing rules for a source text to be translated by RBMT systems. Somers (1997) hinted at this technique by proposing to “post-edit” the source text. Both approaches share the idea of manipulating the source text once the authoring stage has been completed and prior to MT. An added advantage of automatic rewriting, unmentioned in the literature, is the cost-effectiveness of freeing technical writers up from having to comply with a vast array of CL rules designed to target specific MT system weaknesses.

Babych et al. targeted light verb constructions, “*combinations of a 'semantically depleted' verb and its complement*” (ibid: 36), which require a non-literal translation, and for which the coverage is usually not consistent in MT systems. They wrote rules to automatically change light verbs not included in UDIs into synonymous verbs. The authors do not mention the technique used to do the transformations. If we assume, therefore, that simple search and replace (S&R) techniques were used, the rules may have been lexical, that is, each verb would need a rule. Generalisability is minimal.

Another paper discussing automated pre-processing of texts to be translated is the one by Turcato et al. (2000). These authors report efforts to achieve a more accurate machine translation output in a closed-caption context. They report the use of automatic changes to the source as a pre-analysis module in the MT system. The modifications are carried out to normalise the text. These include the expansion of contractions, the recognition, normalisation and annotation of stutters, or the conversion of number words into digits, to mention but a few.

1.5.2 POST-PROCESSING

The methods to improve machine translation quality mentioned in section 1.5.1 focus on manipulating the source text to increase machine-translatability. In a post-processing stage, however, it is the machine translation output that is edited to obtain the desired quality standard. In this section we will briefly describe the traditional human editing of translation output and focus on more recent attempts to

automate this task. Such automation is nowadays performed by either manually crafting or automatically inferring S&R rules, or using statistical methods which avoid linguistic knowledge.

Post-editing (PE) has long been the main approach to improve machine translation output. It has been defined as the “*term used for the correction of machine translation output by human linguists/editors*” (Veale and Way, 1997). Depending on the required end-quality, two main types of PE are distinguished: rapid post-editing and full post-editing (Loffler-Laurian, 1996). In contrast to full PE, where a target text is edited to a high standard of quality, rapid PE only requires changes to respect the TL syntax and lexicon, and to structures that hinder comprehension.

1.5.2.1 AUTOMATIC POST-EDITING

Due to the deterministic nature of RBMT systems, mistakes are recurrent and could even be categorised. Therefore, Schäfer (2003) claims that post-editors easily learn to recognise them and to apply the required changes accordingly. Post-editors work with three texts: the source, the MT output and their own target text. PE is thought to require high cognitive effort and is regarded by professionals as a boring, repetitive and time-consuming task. If the correction process is repetitive, and the mistakes predictable, would there not be a way to automate PE?

The possibilities of automating the task of post-editing have long been proposed. The use of macros and search & replace options offered by word processors was reported by the Pan-American Health Organization (PAHO) (Vasconcellos, 1987). Allen and Hogan (2000) proposed the creation of an automated post-editing (APE) tool. Somers (1997) emphasises the benefits of using this approach but acknowledges the risks of “*under- and overshooting*” (ibid: 202). S&R options in word processors tend to be quite limited, more often than not making it impossible to account for grammatical requirements for the changes, let alone the introduction of any degree of generalisation.

The concept of S&R was further studied by Guzmán (2007, 2008). This author tried to broaden the limitations of word processor’s S&R options by using regular expressions (Regex). Regex provide the tools to describe text strings by “ignoring” the unimportant characters and “identifying” the significant characters. This allows for the creation of more complex S&R rules. For instance, the author reports the possibility of

correcting “*Estos son programas Java-basado que ejecutan en su navegador Web*” as “*Estos son programas basados en Java que ejecutan en su navegador*”. The search rule looks for a word hyphenated by *-basado* and the replace rule substitutes the string by fronting *basado* and adding a plural *-s* to it after checking that the preceding word, *programas*, ends in *-s*.

All the operations are performed at a string level and the possibility of using the source string to disambiguate or find additional anchor points exists (Guzmán, 2008). Such is the case with titles in the technical documentation starting with *-ing* words. They are often badly translated into Spanish as gerunds, with additional determiners or reflexive pronouns. Guzmán reports a number of regex rules which transform gerunds into imperatives and delete additional elements.¹⁶ However, no linguistic knowledge is extracted from the machine translation output before applying the rules. They are crafted based on the surface string characters and, as a result, the rules are very lexicalised. Their cost-effectiveness, nevertheless, appears convincing given that Symantec and VistaTec (an Irish software localisation service provider) have implemented this approach in their post-processing stages.

Apart from the limitations the Regex syntax might pose, one of the disadvantages of using them is the time required to manually identify and code the rules. In an attempt to automate this process, Elming (2006) proposed the use of transformation-based learning (TBL) to automatically extract rules in a post-processing stage. The rules would be learned from a parallel corpus built using the MT output and its post-edited version. In addition, the MT output would be POS-tagged to increase discriminative information.

Elming reports that from a (English-Danish) training corpus of 220,000 words 1,736 rules were extracted. Despite the good performance of these rules, they were very specific to the training data, this being one of the main disadvantages of the method. However, 20% were then applied to unseen data, with a precision of 70% or higher, and were therefore considered “general”. The author claims that the resulting

¹⁶ No accuracy or efficiency of the S&R rules are reported by Guzmán although it is clear that they cannot account for all instances of the patterns discussed. For instance, the rule aiming at transforming an *-ing* word translated as a gerund into an imperative only works for verbs whose infinitives end in *-ar* in Spanish.

rules are very informative and transparent, which helps in the control and understanding of the process.

Other attempts at improving RBMT rules using TBL have been made, for instance, that of George and Japkowicz (2005), who identified and corrected relative pronouns (French-English) with a high accuracy. Moreover, Font Llitjós et al. (2005) also proposed the method of recording post-editors' work, as well as their oral reasoning of the trigger for the change, as a means of modifying or adding new rules to an existing RBMT system.

1.5.2.2 STATISTICAL POST-EDITING

Although it looked as if the streamlining of SMT had divided the research community into empiricists and rationalists (Somers, 2003), the unification of efforts re-emerged when the benefits of the combination of both methods was understood. Knight and Chander (1994) suggested that a good starting point could be to create a corpus with RBMT output and its post-edited counterpart and to apply statistical MT methods as a way to teach a program which changes to make. In this approach, PE would be conceived of as an intralingual translation (Jakobson, 1959). It was not until recently that Simard et al. (2007a, 2007b) revisited this idea and used an SMT system as a post-processing stage for an RBMT system to automatically post-edit the translation. This method is now widespread and called Statistical Post-editing (SPE).

Let us describe the SMT architecture briefly. SMT systems require a training cycle, where all the statistics are calculated, and a search module to perform the actual translation. Two different models must be trained: the translation model and the language model. The aim of the translation model is to create a bilingual dictionary or phrase table where the most probable word/phrase pairings are listed on the basis of statistics taken from the training parallel corpus.

The success of this phrase table depends largely on word alignment, which is not a trivial task. First, a large training data set is required to estimate the pairings (Brown et al. 1990, Manning and Schütze, 2000). To exemplify what "large" means, the research community reported the use of over 10 million sentence pairs during the Fourth Workshop on Statistical Machine Translation (2009). The suitability of corpora used as training data is a topic which has not received much attention, despite its clear effect

on the quality of the translation and phrase tables. Ozdowska and Way (2009) recently conducted an experiment which clearly showed the need to focus on quality data rather than quantity as opposed to what was assumed previously (Zollmann et al. 2008).

Secondly, languages do not have a one-to-one word correspondence and often one word in a language corresponds to zero, two or more words in another. The number of target words that correspond to one source word is called *fertility*. Thirdly, languages do not only differ in the amount of words used to convey the same meaning, but also in the distribution of these words. The difference in word ordering is called *distortion*. If all these characteristics are to be addressed the required algorithm for the creation of the phrase table is quite complex.

The language model is responsible for the training of the system in the TL. The target language can be learnt following several methods. The most popular is the N-gram model (Brown et al. 1990). Basically, this model regards learning as a word prediction task. According to the *Markov assumption* it suffices to know the last few words to predict the next (Manning and Schütze, 2000). Therefore, sequences with a specific number of words, that is, n-grams, n being the number of words in the sequence, are listed according to their probability of occurrence (ibid). It could be argued that the higher the n-gram level, the more fluent the output will be. However, it should be remembered that on the one hand, the repetition rates of long sequences of words are low, and therefore, the probability of the string would be low, and, on the other hand, the amount of calculations required for long strings is enormous, and combinatorial explosion problems may arise (Arnold, 2003).

Other language models include clustering and probabilistic parsing methods. The former is based on the grouping of similar words depending on the distribution of neighbouring words. The latter is based on parsing, that is, on an automatic grammatical analysis of the target sentences for the learning process and reconstruction of sentences for the search model (Manning and Schütze, 2000).

Once the translation language models are trained, it is possible to use the system for the automatic translation of new sentences. First, the most probable words are selected from the translation model dictionary. Then, the statistics stored during the language model training process are consulted to calculate the most probable order in which

these words should appear, and whether any should be deleted or added. It should be noted that when searching for a new translation, it is assumed that the TL and SL follow the same distribution. Allowing for a degree of distortion and fertility, the combination of the best probabilities will produce the output (Somers, 2003).

The architecture described above is a pure SMT system. However, as we saw with RBMT systems, there is now a tendency to combine advantages from both architectures. For SMT systems, this means introducing linguistic knowledge into both the training and translation cycles. As early as 2003, researchers started suggesting the benefits of including syntactical information, either in the form of part-of-speech taggers or more sophisticated parsers (Och, et al. 2004; SSST 2007-2009). The ideas have varied from the use of parse-trees to learn phrasal alignments (Lavie et al. 2008; Nakazawa and Kurohashi, 2009) and language models (Charniak et al. 2003) to introducing reordering models using syntactic information from source text trees (Hashimoto et al. 2009).

In order to build an SPE system, an SMT system is trained using a parallel corpus constructed with raw MT output and its aligned reference translation or post-edited version. The MT output is treated as the SL and the reference as the TL. Then, the trained system is used to ‘translate’ between raw MT output and the post-edited translation. Simard et al.’s experiments proved successful, improving the MT system output by up to 4 BLEU points, and even out-performing a state-of-the-art SMT system used on its own. Interestingly, they reported that there was little variation in score depending on whether the target language corpus to train the system was a human translation or a post-edited version. Because post-edited versions tend to be closer to the MT output structurally and lexically, systems - and metrics - developed on string-based approaches have been shown to perform better when compared against post-edited output (Snover et al. 2006).

Although claims have been made about quality improvements by different groups using SPE and increases in BLEU scores reported (Simard et al. 2007a, 2007b; Dugast et al. 2007, 2009; Isabelle et al. 2007; Schwenk et al. 2009), the actual changes made have not been studied. As with most corpus-based learning methods, it is not possible to generalise in linguistic terms the changes made by the SMT module, that is, one cannot say that the SPE module targeted and corrected the overuse of reflexive

pronouns or incorrect choice of a particular conjunction. One can either qualitatively describe the type of post-editing change made by the SPE module (Dugast et al. 2007), or use a grammar-independent error classification (Vilar et al. 2006; Font Llitjós et al. 2005). The former follows the traditional grammatical feature error description – whether lexical or structural. The latter is quite rooted in the word-level and tries to capture the transformation errors post-SPE texts can have, e.g. missing words, word order or incorrect words. Whereas a combination of both proves a good framework for describing the performance of the SPE system and the quality of its output (Tatsumi and Sun, 2008), it is still not clear how to proceed to fix specific error types. Needless to say, it is always possible to do a raw count of all the changes made by the SPE module and report the number of ‘improvements’, ‘degradations’ and ‘equivalent effect’ (Dugast et al. 2007).

1.6 CHAPTER SUMMARY

Chapter 1 introduced the core domains pertaining to this research, i.e. the nature of -ing words and how they can be ambiguous for both humans and MT; the current MT paradigms in use, with a focus on RBMT, which is the system type used in this study; MT evaluation techniques, including human and automatic metrics; and the pre-processing (CL and automatic re-writing) and post-processing techniques (automated post-editing, including SPE) available for improving SL input and TL output from MT systems.

This prepares the ground for Chapter 2, which attempts to set up a comprehensive categorisation of -ing words. This categorisation is essential for a complete quantification and evaluation of the issues involved in the translation of -ing words, as well as the solutions that are most appropriate for each of those issues.

CHAPTER 2

CHAPTER 2: EVALUATION METHODOLOGY

The aim of this Chapter is to discuss the theoretical and practical methodological frameworks used to perform the evaluation of -ing words. The first objective of this dissertation is to identify which -ing words within the IT-domain procedural and descriptive texts are problematic for the RBMT system. The first step towards this goal is to collect a sample of representative data. Section 1 is concerned with considerations for corpus design and the practical decisions made regarding the data extraction. Evaluation approaches are examined in section 2 and the practical reasoning for our human binary attribute evaluation explained. Section 3 explores automatic evaluation and sets the framework for comparison against the human evaluation results.

Let us here define the concept of procedural and descriptive texts. Byrne (2006) suggests that technical publications can be categorised as procedural documents, descriptive and explanatory documents, persuasive or evaluative documents and investigative documents. User guides and installation manuals belong to the first and second categories. In fact, technical documentation is written based on topics, specifically three types: descriptive, procedural, and reference. Descriptive topics describe concepts or products. Procedural topics instruct on how to achieve a particular objective, be it installing a software program or performing a particular action a product allows. Finally, reference topics direct the user for cross-referencing. Given that our corpus comprises user guides and manuals we considered it appropriate to restrict the generalisation of our findings to IT-domain procedural and descriptive texts.

Throughout the Chapter, the principles of research design described by Frey et al. (1991) are drawn upon to ensure external and measurement validity, generalisability and to avoid operationalisation flaws.

2.1 BUILDING A CORPUS

2.1.1 CORPUS-BASED APPROACH

The collection of sentences containing -ing words to be machine translated and judged by evaluators could be accomplished in two different ways: a corpus or a test-suite. A corpus is “*a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria*” (Bowker and Pearson, 2002: 9). The use of

corpora for linguistic research has been widely approved due to their capacity to incorporate the exact occurrences which can be found in real-life scenarios (Biber, 1988; McEnery and Wilson, 1996; McEnery et al. 2006). On the other hand, a test suite has been defined as "*a collection of (usually) artificially constructed inputs, where each input is designed to probe a system's treatment of a specific phenomenon or set of phenomena*" (Balkan et al. 1994: 24). Supporters argue that test suites provide the opportunity to isolate the linguistic structures of study and perform an exhaustive analysis of all the possible combinations, grammatical and ungrammatical, a specific linguistic phenomenon offers, with the certainty that each variation will only appear once (Balkan et al. 1994).

In order to decide which of the approaches best suited our research, our particular requirements and goals were further analysed. As mentioned above, our aim is to identify -ing words in IT-domain procedural and descriptive texts which are problematic for the RBMT system. The first thing to verify is whether using purposely crafted sentences or real-life sentences would make any difference. The question that therefore emerges is: Do IT texts have domain-specific linguistic features which can affect the use of -ing words? Calsamiglia and Tusón (1999) mention that one of the main concerns of text scholars has been that of classifying texts. Many classifications have been proposed following different criteria. It is the linguistic criteria, explored by Biber (1985, 1986, 1989) which concerns us most. Should linguistic features and their occurrence frequencies vary from text to text depending on the genre, register or domain, it would be essential to analyse the -ing words in the particular contexts in which they appear. Biber (1988, 2006) and Biber et al. (1998) and Reppen et al. (2002) claim linguistic variability in different registers, that is, "*the varieties of language that we use in different situations*" (Biber et al. 1998: 2), reporting varying syntactic patterns and occurrence frequencies for the same linguistic phenomena. Biber carried out a detailed study which counted the occurrence frequencies of a series of linguistic features in different genres (1988). Among others, he included three different subgroups of -ing words: gerunds, present participials functioning as reduced relatives and present participial clauses. For gerunds, for instance, he reported an occurrence rate of 10.6/1,000 and 6.5/1,000 words in official documents and general fiction respectively. Present participle reduced relative clauses occurred 4/1,000 words in

official documents and 0.7/1,000 words in general fiction. Present participial clauses, in turn, occurred 0.3/1,000 words in official documents and 2.7/1,000 words for general fiction. These numbers show that -ing words can appear in varying occurrence rates across genres. RBMT systems rely on a grammatical and lexical analysis of the source to generate the translation. Therefore, the exact patterns and lexical items determine the output. This brought to the fore the importance of using real data.

It was essential for this research to identify and assess the actual uses of -ing words specific to IT-domain procedural and descriptive texts. No study has listed the specific patterns and occurrence rates of -ing words in IT domain texts. Although the use of exhaustive test suites or a combination of test suites and corpora has been proposed by many authors (Heid and Hildenbrand, 1991; Lepage, 1991; Arnold et al. 1994; Balkan, 1994), they are mostly preferred for what EAGLES describes as a *diagnostic test* (Balkan et al. 1994). This type of test is performed by developers to test the overall performance of a system with the aim of identifying deficiencies, for which a controlled set of linguistic features and their combinations is needed. A corpus, on the other hand, would allow us to focus on authentic IT texts where -ing words occur according to the writing norms specific to the genre (McEnery et al. 2006: 6-7). By using the appropriate design principles and analysis tools reliable, quantitative data could be yielded. As a result, the option of building a test suite was discarded.

2.1.1.1 CORPUS DESIGN

The use of corpora for linguistic studies is widely accepted, confirming its validity through the *panel* approach (Frey et al. 1991). A corpus is often used to reduce evaluation/analysis effort while maintaining the option to extrapolate the results to the original groups for which the research question has been proposed (the population). Therefore, it is essential to ensure its external validity. The literature in corpus design warns of the challenges of guaranteeing ecological and sample validity and calls attention to points that should be carefully addressed when building a corpus. Authors concur that the decisions made must depend on the purpose and resources of each study (Kennedy, 1998: 60-85; Bowker and Pearson, 2002: 45-57; Olohan, 2004: 45-61). McEnery et al. (2006) agree with Atkins et al. (1992: 6), who comment that: “*It would be short-sighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as*

‘unreliable’ or ‘irrelevant’ because the corpus used cannot be proved to be ‘balanced’.

Kennedy (1998: 60-70) points out three design issues to take into consideration when building a corpus: stasis and dynamism, representativeness and balance, and size. The better these three factors are handled in a research project, the stronger external validity it will hold. A dynamic corpus is one that is constantly upgraded whereas a static corpus includes a fixed set of texts. The aim of the present research is to study the current performance of our RBMT system when dealing with -ing words given the current level of system development and source text quality. Dynamic corpora are mainly used when trying to capture the latest uses of language or when studying linguistic changes over time. It would have been convenient to use a dynamic corpus if the use of -ing words varied as time passed. However, a change in the use of grammatical structures for a particular text type is not expected to be significant in a short period and therefore, it was decided that the use of a static corpus would be satisfactory.

Representativeness is the second design issue discussed by Kennedy (ibid: 62-65). He points out that it is difficult to ensure that the conclusions drawn from the analysis of a particular corpus can be extrapolated to the language or genre studied (ibid: 60). Along the same lines, Balkan et al (1994: 18) warn that even in the case where one might be interested in the combinations of a linguistic pattern that occur “*in real life*” only, one should remember that corpora only show a “*more or less representative sample*”. To guard against the corpus violating external validity, additional control points outlined by Bowker and Pearson (2002: 49-52) were taken into account when selecting the texts: text length, number, medium, subject, type, authorship, language and publication date. The latter argue that texts that vary greatly in length might use different linguistic features due to space constraints. However, the difference in length in IT-domain texts depends on the complexity of the product. The format of the guidelines, and subject and text type remain the same. We argue therefore that the concentration and patterns of -ing words should also remain constant despite text length. Moreover, the texts are written by various teams composed of several writers, which minimises the possibility of the results being affected by the idiosyncratic uses of -ing words. The texts provided by Symantec included different types of guides, from

Installation Guides to Getting Started Guides or Maintenance Guides for administrators, and belonged to three different products. Full texts were used to ensure -ing words from all sections were included, should they be different depending on their location in the text (e.g. titles, introductory descriptions or bulleted lists). Electronic versions of the texts were made available in FrameMaker and XML formats, written in English, as this is the language used for development across the company. Note that we decided to use texts which had not undergone any language control, that is, the selected texts were not written following the controlled language guidelines nor were they approved by the controlled language checker in operation at Symantec. This would make it possible to measure the extent to which -ing words cause problems prior to the application of CL rules, and to what degree their control is necessary.

According to McEnery et al. (2006), in order to obtain a high representativeness in a specialised language corpus, as in our case, fewer resources than in a corpus for natural language might be necessary. This claim arises from the assumption that a sublanguage is “*a version of a natural language which does not display all of the creativity of that natural language*” (McEnery and Wilson, 1996: 148) and therefore tends to “*closure*” (ibid: 166) or “*saturation*” (Belica, 1996: 61-74 in McEnery et al. 2006: 16), that is, “*the feature appears to be finite or is subject to very limited variation beyond a certain point.*” To date, the measurement of closure has focused primarily on lexical features. However, McEnery and Wilson tried to adapt the measurement to a syntactic level and analysed sentence types. The two general language corpora they used showed a very low repetition rate of sentence types (6.28% to 2.17%). In contrast, the IBM corpus showed a 39.8% repetition rate. This sets the ground to suggest that sublanguages avail of restricted grammatical variation. It also showed that smaller corpora might suffice to study grammatical features.

The discussion of representativeness is closely related to the third issue mentioned by Kennedy: corpus size (Kennedy 1998: 66-69). Kennedy makes it clear that the size of the corpus should depend on the linguistic phenomena of study (ibid). Bowker and Pearson (2002: 48) go a step further and claim that in studies related to language for specialised purposes (LSP) corpus sizes ranging between ten thousand to several hundreds of thousands of words have proven “*exceptionally useful*”. Our final corpus included around 494,618 word occurrences, which falls within the proposed range.

We carefully considered the ecological validity of our corpus in order to ensure maximum external validity. All issues pointed out in the literature were reviewed and compared against our research requirements. By using real texts that meet the relevant number and authorship variation, and stability in subject, type and medium as required for the population for which we intend to draw conclusions, we feel that the ecological validity of the corpus was ensured.

2.1.2 PILOT STUDY

As Frey et al. (1991) suggest, one of the better practices to minimise measurement error and therefore increase reliability is to carry out a pilot study. Therefore, once the corpus was compiled, an exploratory study for the evaluation process was performed. One pattern where -ing words frequently appeared was chosen for the pilot: preposition or subordinate conjunction followed by an -ing word.¹⁷ The corpus was searched for sentences containing this pattern and 1,857 instances were found. Next, these sentences were machine translated using SYSTRAN into French, German, Japanese and Spanish and evaluated by one evaluator per language. The evaluators were translators or MA students in Translation Studies with experience in MT who were native speakers of the target languages.

The evaluation included the source sentence and raw machine translation. It sought to learn whether the structure “preposition/subordinate conjunction + -ing” was translated grammatically, comprehensibly, and accurately.¹⁸ Evaluators were asked to focus on the structure of study only, instead of taking the whole sentence into account. We are aware of the fact that different linguistic phenomena could affect different parts of a sentence and affect one another when using an MT system. However, due to the high complexity involved in tracking the impacts, particularly when no information about the MT rules is available, we opted for the feature-based approach. In addition, in order to obtain specific results, evaluators were also asked to indicate whether the problem was due to the preposition or subordinate conjunction, or the -ing word itself.

An additional question addressing style was introduced where evaluators were asked whether the examples needed post-editing or not. The main goal of this question

¹⁷ acrolinx IQ uses the PennTreebank tagset. This set assigns the tag IN to both prepositions and subordinate conjunctions. Hence the inclusion of these two categories under the same pattern.

¹⁸ See 2.2.1.4 for a discussion on the attributes selected for evaluation.

was to prevent evaluators from feeling tempted to classify an -ing example as incorrect because they thought the sentence in general should not be published without being post-edited, despite being grammatical.

Finally, a fourth question which aimed at identifying whether the incorrect translation of an -ing word was due to difficulties in the analysis or the transfer/generation stage of MT was included. Scholars argue that trained translators develop language, textual, subject, cultural and transfer competences (Neubert, 2000:5-10). The grammatical awareness required to translate different speech levels is also emphasised by Mailhac (2000). However, it is only in recent years that translator training programmes have started to include some introductory modules on MT systems. Therefore, in reference to the abovementioned questions, it was clear that translators would be able to decide whether a sentence in their native language was grammatical, comprehensible and conveyed the same meaning as the original. However, it was not clear to what extent the evaluators, despite their experience with MT, would be able to attribute problems either to MT source sentence analysis or transfer/generation. It was finally decided to maintain the question to see how well MT-specific questions were handled by translators and MA students.

Counts per language showed that 70-80 % of the isolated -ing words had been classified as correct. Also, specific subgroups within the studied structure seemed to be more problematic across languages. In fact, 6 out of the 33 subgroups accounted for 65% of the -ing words classified as incorrect (see Table 2.1).

		Spanish		French		German		Japanese	
Preposition	Examples	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
by + ing	377	351	26	358	19	364	13	301	76
for + ing	339	243	96	284	55	262	77	224	115
when + ing	256	205	51	2	254	213	43	161	95
before + ing	163	145	18	146	17	145	18	134	29
after + ing	122	107	15	117	5	114	8	108	14
about + ing	96	82	14	82	14	88	8	88	8
on + ing	89	38	51	80	9	58	31	29	60
without + ing	75	47	28	65	10	71	4	66	9
of + ing	71	65	6	65	6	60	11	57	14
from + ing	68	30	38	31	37	24	44	33	35
while + ing	54	3	51	45	9	27	27	44	10
in + ing	36	27	9	9	27	23	13	9	27
if + ing	19	15	4	10	9	17	2	17	2
rather than + ing	14	0	14	0	14	0	14	1	13
such as + ing	13	9	4	9	4	9	4	8	5
others	65	26	39	38	27	39	26	23	42
TOTAL	1,857	1,393	464	1,341	516	1,514	343	1,303	554
%		75.01%	24.99%	72.21%	27.79%	81.53%	18.47%	70.17%	29.83%

Table 2.1: Results for correctness across languages (pilot project)

Through the pilot study we identified a few points that needed further consideration and refinement. In the first place, a fuller picture of the different contexts where -ing words occur was necessary in order to draw an overall conclusion on this linguistic feature. Most importantly, we noticed that despite being included under the same pattern - preposition or subordinate conjunction followed by -ing word - the relation between the two parts of the structure could be very different. To mention but one example, an -ing word was sometimes mandatory after a prepositional verb, e.g. *To prevent Backup Exec from communicating with a Web server, clear the check box next to that Web server's name*, whereas in others, the -ing word was used in a temporal adverbial clause to avoid repeating the subject, e.g. *Include previously discovered resources when sending notification*. The issue with grouping such different linguistic phenomena under the same heading is that they may not necessarily be translated following the same patterns in the target language. Therefore, the need for a more appropriate classification of -ing words arose. This would increase content validity and help interpret the results of the evaluation appropriately (see section 2.1.3).

Secondly, evaluators were asked to focus on a particular -ing word and its translation. However, delimiting the exact target language words that needed to be evaluated was reported not to be straightforward, especially for Japanese, where sentence structure is disparate and ideas are formulated differently. Another difficulty was, as mentioned above, that because different linguistic phenomena could be entangled in the structure, doubts could arise as to whether the translation of the -ing word should in fact be rated as “correct”.

Finally, the reliability of human evaluation must be mentioned. Human evaluation is often accused of being subjective and inconsistent. The internal validity of a study can be compromised by these threats, so careful consideration was given to this challenge. Firstly, the number of evaluators would have to be increased. Secondly, we noted that for the question where the evaluators were asked to judge whether the incorrect output was due to a source language analysis ambiguity or not, the evaluators were not always able to respond correctly. Therefore, in addition to discarding this question, it was decided that only professional translators with experience in machine translation/post-editing would be used. Also, the risk of the poor quality of a sentence influencing the decision on the correctness of the translation of an -ing word was considered. It was decided that measures would be taken to increase overall sentence quality and an acceptable reference translation would be provided as a guideline for evaluators.

2.1.3 CLASSIFICATION

It was clear from the pilot study that -ing words would need to be classified on a more fine-grained level. Depending on the interaction of the -ing word with its context, the translation requirements changed. It was thought that the context could serve as an indicator of the specific translation requirements and therefore, should be considered for inclusion in the classification.

When deciding about the classification approach to be taken, the option of reusing an existing classification or proposing a study-specific scheme was considered. Sampson (2002) claims that the key to a systematic linguistic analysis is a detailed and consistent classification schedule as “[n]o empirical science is likely to be able to advance very far without a set of recognized standards for classifying and recording

data” (ibid: 74). Sampson regrets that taxonomy does not attract much attention in the field of computational linguistics, where each research centre tends to create its own standards and classification. A common classification scheme allows results to be compared and understood by the whole research community. Still, he acknowledges that sharing a scheme might not be straightforward due to disagreement in class boundaries (ibid). To illustrate the gravity of the situation, he reports the inability of nine research groups participating in a workshop (Association of Computational Linguistics at Berkeley California, 1991) to reach an agreement on the manual parse of a sentence. *One of those capital-gains ventures, in fact, has saddled him with Gore Court* was to be parsed and only two clauses, [with[Gore Court]], were agreed upon! Following Sampson’s advice, existing classifications were reviewed with the aim of selecting the one that was most appropriate for our study.

2.1.3.1 FUNCTIONAL CLASSIFICATION

As mentioned in Chapter 1, words ending in -ing have been divided into gerunds and participles (Quirk et al. 1985). However, contemporary grammarians claim that this division can no longer be defended and propose that words with a verb base and the -ing suffix be classified as gerundial nouns (genuine nouns), gerund-participles (forms with a strong verbal flavour), and participial adjectives (genuine adjectives) (Huddleston and Pullum, 2002).

Classifying -ing words as gerundial nouns and participial adjectives would provide us with information about their behaviour (they function as genuine nouns and adjectives) and they could be more easily mapped to the target language renderings. In contrast, gerund-participles appear in a wider range of structures which results in several translation options depending on the particular structure and the target language. Consequently, a deeper classification of -ing words was considered necessary.

In grammar books, (Quirk et al. 1985; Carter and McCarthy, 2006; Biber et al. 2002) -ing words are thoroughly described under the sections of different types of word classes, phrases or clauses in which they can appear, that is, a syntactic description of the -ing word is spread throughout the books. However, no classification has focused on -ing words as the main topic and in a detailed manner.

Izquierdo (2006) faced this deficiency when carrying out a contrastive study of the -ing form and its translation into Spanish. She proceeded by gathering the syntactic and functional description of the literature published on the linguistic feature.¹⁹ She then compiled a parallel corpus and analysed the -ing words, comparing the theoretical framework previously established from the readings and the actual uses found in her corpus.²⁰ As a result, she established a functional classification of -ing words (see Table 2.2). Other works have been published on -ing words (Doherty, 1999; Behrens, 1999; Römer, 2005; Espunya, 2007). They tend to focus on a particular realisation of -ing words. For instance, Doherty focuses on adverbial free -ing adjuncts, Behrens and Espunya examined free -ing-participial adjuncts and Römer performed a comprehensive study of the oral use of progressives.

Functions	Adverbial	Progressive	Characterisation		Referential
Grammatical structures	-time	-past	pre-modifiers	post-modifiers	-catenative
	-process	-present	-participial adjective	-reduced relative clause	-prepositional clause
	-purpose	-future		-nominal adjunct	-subject
	-contrast	-conditional		-adjectival adjunct	-direct object
	-place	-etc.			-attribute complement
	-condition				-comparative subordinate
	-etc.				

Table 2.2: Summary of the main functional categories of -ing forms observed in the study of Izquierdo (2006)

Izquierdo's classification, which unifies dispersed works, was considered suitable for our study for several reasons. -ing words cannot be classified in isolation; contextual information must be considered in order to distinguish a gerundial noun from a participial adjective or a gerund-participle. The functional classification would

¹⁹ Consulted grammars: Quirck et al. 1985; Biber et al. 1999; Huddleston and Pullum, 2002. Other consulted works: Duffley, 2005, on the gerund-participle; Wierzbicka, 1988; Römer, 2005, on progressives; Leech, 1987; Rabadán et al. 2006; Emonds, 1985; Brinton, 2000 on the characterisation function -ing constituents; Bäcklund, 1984, on adverbial -ing constituents; Kortmann, 1991, on adjuncts; Jespersen, 1954, on adverbial -ing constituents; Ramón García, 2003, on characterisation of -ing constituents.

²⁰ ACTRES Parallel Corpus: size 2 million words. Divided into five subcategories: fiction, non fiction, newspaper, magazines and miscellaneous. Texts mostly dating from 2000-2006.

provide the criteria to establish this context, as well as being an adequate framework for discussions with technical writers, who are not necessarily linguists, nor do they have experience in MT. Moreover, the classification categories identify precise grammatical structures where -ing words occur. Given that our MT system is a rule-based system, a deterministic output based on source grammatical structures – and lexicon – is expected, facilitating the analysis and identification of issues. Additionally, the grammatical structures in which -ing words occur can be described by means of syntactic, POS or morphological characteristics. Patterns written based on this information could be used to (semi-)automatically assign each -ing word to a classification category.

Although Izquierdo's classification offered all the abovementioned advantages, there is an observation to make on its suitability. The corpus she used to establish the classification was compiled from general language, mainly fiction texts, whereas our corpus is IT-domain specific, covering procedural and descriptive text types. The difference in subject-domain should not concern us excessively, as the main difference between the texts would lie in terminology and we focus on a grammatical feature. The fact that the text types do not map directly raises some concern. Biber defines text types as "*classes of texts that are grouped on the basis of similarities in linguistic form, irrespective of their genre classification*" (Biber, 1988: 206). This means that text types share the particular usages of linguistic features, which occur regardless of the genre. In order to ensure that Izquierdo's classification held for our research, recall for the classified -ing words was measured. 80% of the -ing words in our corpus could be classified using Izquierdo's scheme. We will discuss the remaining 20% under 2.1.3.3.

2.1.3.2 -ING CORPUS

Once the corpus was built, we identified all sentences which included an -ing word and created our -ing corpus. This task was done automatically by programming the CL checker to identify and list all sentences including -ing words. An -ing word was described as a word ending in "-ing" which could take a JJ (adjective), NN (noun) or VBG (gerund) as POS tag. In addition, a stop-list was included to avoid irrelevant words such as *string*, *bring* or *thing*, for instance, from being included. The CL checker returned sentences for 10,417 -ing instances. Our 2.1% concentration falls close to the 1.5% reported by Biber (1988) for official documents and the 1.02% of the BNC

written instructional texts (4,489 -ing forms out of 440,548 words) which was reassuring from a generalisability point of view.²¹

2.1.3.3 EXTRACTION

As mentioned in section 2.1.3.1, instead of describing and classifying all the -ing words in the corpus manually one by one, patterns based on syntactic, POS and morphologic information were used to automate the allocation. The CL checker was programmed to search for the specific patterns. We ensured that the rule descriptions were as inclusive as possible. On a first pass, the linguistic structures were described as broadly as possible. This returned all the sentences with the relevant linguistic structure and other irrelevant structures which happened to fit the pattern. On a second pass, the groups of sentences returned by the checker were manually cleaned. For instance, post-modifying -ing words occurred, among others, in reduced relative clauses. A search for reduced relative clauses was performed by writing a rule which described them as a word tagged as a noun followed by a word ending in -ing (see Figure 2.1).

reduced relative clauses PATTERN noun + ing RULE find: noun (adverb)* ing word Highlight: noun ing word
--

Figure 2.1: Model of the rule used to extract reduced relative clauses.

A noun could be singular, plural, common or proper. An -ing word could be a noun, adjective or a gerund/participle ending in -ing. The presence of adverbs between the noun and the -ing word was allowed. In the rules, the patterns to be searched are described at a string-level, composed of a number of objects. Then, the objects to be highlighted are specified.

This rule would detect reduced relative clauses. However, many instances where a noun is followed by an -ing word but the structure is not a reduced relative clause would also be detected (see Table 2.3). The irrelevant instances were discarded during the cleaning cycle.

²¹ Note that the concentration of linguistic features in a corpus usually ranges around 0-10% (Biber, 1988). The frequency rates for other potentially problematic features for RBMT system in our corpus were 1.47% for passives, 2.3% for coordinating conjunctions or 0.35% for 3rd person pronouns, for instance. A 2.1% frequency rate for -ing words, therefore, is a considerable concentration to be studied.

Relevant	Reduced relative clause: Place the media containing the data you want to restore in the storage device.
Irrelevant	Adverbial clause (manner): We recommend that you install bv-Control for UNIX using install.sh

Table 2.3: Example of relevant and irrelevant matches of the rule written to search for reduced relative clauses

Despite the need for a cleaning cycle, this procedure was considered more focused, quicker to perform and less error-prone than a completely manual classification. When the extraction was completed, 8,316 examples were classified out of a total of 10,417, that is, around 80%. As mentioned in section 2.1.3.1, the high recall obtained ensures the validity of using Izquierdo's classification. In order to get an overview of the -ing words not extracted semi-automatically, a representative random sample was obtained from the remaining 20%. 329 examples were manually classified. The majority of the -ing words were pre-modifiers (44%) out of which 30% included the -ing word *following* (see Table 2.4 for a summary). We observed that all -ing words were already represented in the 80% classified. -ing words functioning as adverbials of mode not introduced by any preposition accounted for 16% of the examples. We noted that in all cases the -ing word was *using*. A number of examples were already included in the 80% classified. Gerundial nouns accounted for 10% of examples as did gerund-participles in subject function. Whereas the classified 80% contained a number of gerundial nouns, no -ing words functioning as subjects were included in our classification. This subcategory of referential function is referred to by Izquierdo. However, due to the lack of a suitable pattern to extract them automatically, we did not address them. Adverbials of elaboration accounted for 5%. These cases are also -ing words without a differentiating pattern and a reduced number was included in the 80% classified during the cleaning cycle. The remaining 15% of the examples were spread across over 20 subcategories for which examples exist in our 80% classified. No new subcategories appeared. Therefore, we concluded that checker precision might have caused these cases to be overlooked.

Category	Subcategory	Frequency Rate	Examples
Characteriser	pre-modifier	44%	Desktop Agent low disk warning threshold
Adverbial	mode	16%	Roll forward using logs
Referential	gerundial noun	10%	Concurrent processing .
Referential	subject	10%	Typing < host name > only works with properly configured DNS.
Adverbial	elaboration	5%	Type the path to the folder, including the folder name.

Table 2.4: Examples of the 20% -ing words not found during the semi-automatic extraction

BREAKDOWN OF CLASSIFICATION AND EXTRACTION²²

Rules were written to identify structures pertaining to subcategories of Izquierdo's Adverbial, Progressive, Characteriser and Referential categories. Additionally, a category to include Titles was added, given their occurrence rate in IT documentation.²³ In fact, this category was the most populated with 2,603 instances. Characterisers and Adverbials also included high numbers of instances with 2,488 and 1,970 respectively. Progressives and Referentials counted fewer instances with 661 and 594 respectively.

Characterisers

As the word suggests, characterisers modify or provide information about the noun or adjective they modify. Ramón García (2003) claims that despite -ing words being a means for characterisation, real usage shows that adjectives and relative clauses, for instance, are more frequently used to fulfil this function. Depending on their position regarding the head they modified, Izquierdo distinguished pre-modifiers and post-modifiers. Pre-modifiers can be either adjectives or nouns which are placed before the head of the noun phrase they modify. The function of adjectives is that of modifiers. Nouns, however, have an inherent referential nature. Sometimes, however, nouns are also used as modifiers and, in such cases, function as adjectives.

Desktop Agent low disk warning threshold (nucleus)
Troubleshooting tips (nucleus) for true image restore

Table 2.5: Examples of pre-modifiers

²² For a full classification table see Appendix A.

²³ Note that whereas the classification presented by Izquierdo 2006 consisted of four categories - Adverbials, Progressives, Characterisers and Referentials - the final classification presented by Izquierdo 2008 was extended to include a fifth category of Titles.

There is disagreement over whether noun modifiers should be considered as nouns or adjectives (Biber et al. 1998; Jurafsky and Martin, 2009). The -ing words that are mostly used as adjectives will be perceived by humans as such, whereas those that tend to be used as nouns will tend to be perceived as nouns.

<p>Overwriting existing backup files.</p> <p>Make sure that the cleaning tape is located in the slot that you defined as the cleaning slot.</p>

Table 2.6: Examples of pre-modifiers with a predominant adjectival and verbal flavour

However, given that our classification is functional and that the RBMT system will have to translate them as modifiers, we decided not to further divide pre-modifying -ing words into gerundial nouns or participial adjectives. In total, 1,873 examples were classified as pre-modifiers.

Post-modifiers present three different syntactic structures: reduced relative clauses, nominal adjuncts and adjectival adjuncts (see Table 2.7). Reduced relative clauses (Quirk et al. 1985) are relative clauses which follow the noun they characterise directly, without the need of a relative pronoun or verb, i.e. *who is*, *which are*. When searching the corpus, reduced relatives were described as words with a noun POS tag followed by an -ing word. 377 examples were classified.

<p>To create an XML file (noun) containing (-ing word) all parameters, use the /XML:</p>

Table 2.7: Example of a reduced relative clause introduced by an -ing word

Nominal adjuncts also modify the preceding noun. However, in these cases, the -ing words are introduced by a preposition governed by the characterised noun. When searching the corpus, a nominal adjunct was described as a noun followed by a preposition followed by an -ing word. In total 226 examples were assigned to this subcategory.

<p>It can also guide you through process (noun) of (preposition) creating (-ing word) a non-bootable disaster recovery CD image.</p>

Table 2.8: Example of nominal adjunct

-ing words can also appear as adjectival adjuncts. These characterise adjectives instead of nouns, and, once again, the -ing word is introduced by a preposition governed by the adjective that is being characterised. When searching the corpus, these adjuncts were described as an adjective followed by a preposition followed by an -ing word. 12 examples were included in this subcategory.

We hope that you have found this document useful (adjective) in (preposition) **learning** (-ing word) more about the product.

Table 2.9: Example of an adjectival adjunct

Overall, categorisers accounted for 23.88% of the total number of -ing words in the corpus with 2,488 examples.

Characterisers			
Position	Type	Nº of examples	Examples
Pre-modifiers	modifiers (participial adjectives and gerundial nouns)	1,873	Therefore, irrespective of the routing configuration, the correct IP address is always communicated to the Information Server.
Post-modifiers	reduced relative clauses	377	To create an XML file containing all parameters, use the /XML:
	nominal adjuncts	226	It can also guide you through process of creating a non-bootable disaster recovery CD image.
	adjectival adjuncts	12	We hope that you have found this document useful in learning more about the product.
TOTAL		2,488	

Table 2.10: breakdown of -ing words found in noun or adjective modifying function

Adverbials

-ing words with adverbial function were the second group described by Izquierdo. As she pointed out, often -ing words are preceded by a subordinate conjunction or preposition, making the semantic classification more obvious, but they can also appear on their own. Due to the difficulty of automatically searching the latter, that is, -ing words not preceded by subordinate conjunctions or prepositions, we mainly focused on the adverbial clauses introduced by such conjunctions. Their description was as follows: a preposition or subordinate conjunction directly followed by an -ing word.²⁴ It should be noted however, that when adverbial clauses directly introduced by -ing words were found while searching for other patterns, the examples were transferred to this subcategory.

²⁴ The possibility of adverbs being placed between the preposition/subordinate conjunction and the -ing word was also considered. Also, note that the Penn Treebank tagset does not distinguish between prepositions and subordinate conjunctions and, therefore, the checker would automatically look for both categories.

<p>After running the scripts, run the backup job again.</p> <p>This is done by configuring blackout windows.</p> <p>A failure occurred querying the Writer status.</p>

Table 2.11: Examples of adverbial clauses with -ing words as heads

It is worth noting that classifying a clause introduced by a preposition either as an adverbial clause or as an adjunct is not always a straightforward task (see Table 2.12). A difficulty appeared with the structure *for* + *ing* because the -ing word could be interpreted as the modifier of the preceding noun phrase or an adverbial clause describing a purpose. It was decided that because the texts in the corpus are instruction manuals and therefore mainly procedural, ambiguous cases would be classified as adverbials.

<p>Configure a target SQL Server for auditing.</p> <p>Specifies rules for handling the template job start times.</p>
--

Table 2.12: Examples of ambiguous *for* + -ing structures

Adverbial clauses accounted for 19% of the total -ings in the corpus with 1,970 examples. They presented a variety of introductory prepositions and subordinate conjunctions (see Table 2.13).

Adverbials			
Clause type	Prep. or sub. conj.	Nº of examples	Examples
Manner	By	516	This is done by configuring blackout windows.
	Free	159	Uninstalling Backup Exec using the command line
	Without	88	Change physical tape devices without providing resource credentials.
Time	When	313	When creating the group, use lowercase letters.
	Before	179	Prompt before overwriting imported media
	After	139	After running the scripts, run the backup job again.
	While	65	Job failed while being dispatched.
	On	8	The UNIX registration service adds the target information to the database of the Information Server on executing the setup.sh.
	Through	5	The IDR Configuration Wizard guides you through setting an alternate data path for the *.
	Between	4	Minimum time between closing a job log and starting a new one
	Free	3	A failure occurred querying the Writer status.
	Along with	2	Along with backing up the SAP database files, you should do the following:
	During	2	During cataloguing , Backup Exec reports file formats that it can read.

	From	2	Select this option to have Backup Exec run the tape in the drive from beginning to end at a fast speed, which helps the tape wind evenly and run more smoothly past the tape drive heads.
	In	2	In creating a script file, you would not want to include all entries.
	In the middle of	2	If failover occurs in the middle of backing up a resource, the media that was being used at the time of the failover is left unappendable and new media will be requested upon restart.
	Prior	1	The Emergency Restore feature can be used to restore data for a deleted user if the user data can be restored from a backup of the File Server and a Recovery Password was established prior (sic) making the backup.
	Upon	1	Upon upgrading the Software, all copies of the prior version must be destroyed;
Purpose	For	443	The type of job submitted for processing .
	In	1	This data can be used in planning for additional device allocation, archiving historical data, and improving performance.
Condition	If	20	If cleaning is not possible, the virus is quarantined.
Contrast	Instead of	11	Copying jobs instead of delegating jobs
Cause	Because	2	Because restoring objects from tape requires the creation of a staging location, restoring from tape requires more time than if you are restoring from disk.
Concession	Besides	1	Besides terminating the bus, Y-cables and tralink connectors also allow you to isolate the devices from the shared bus without affecting the bus termination.
Place	Where	1	The following are examples where using true image restore back up files that would not otherwise be backed up:
TOTAL		1,970	

Table 2.13: Breakdown of -ing words found introducing adverbial clauses directly or preceded by a preposition/subordinate conjunction

Progressives

The continuous aspect is expressed in English mainly through -ing words (Izquierdo, 2006). Continuous tenses are constructed by the auxiliary verb *to be* followed by the verb in its -ing form. The auxiliary verb is conjugated, providing information about time and person, whereas the -ing word carries out the meaning of the action that is taking place. In the pattern to search for progressives, they were described as the auxiliary verb *to be* in any of its forms (present, future, past, etc.) directly followed by an -ing word. Additionally, subgroups were created depending on the tense of the auxiliary verb, and where relevant, active and passive voices distinguished.

Specify whether server sets or document sets are being restored. The cluster service should not be running .

Table 2.14: Examples of -ing words with a progressive aspect function

There are instances where participial adjectives functioning as predicative complements and gerund-participles that combine with the auxiliary verb *to be* to construct the verbal progressive aspect are not distinguishable. In cases of ambiguity, the -ing words were considered as part of the verbal periphrasis.

If any prior catalogs are missing , the restore view cannot be expanded. If any files or filegroups are missing , run a Log - No Truncate backup.
--

Table 2.15: Examples of ambiguous instances for progressive-aspect -ing words

-ing words introducing progressive aspect accounted for 6.35% of the total -ing words in the corpus with 661 examples.

Progressives			
Time	Voice	Nº examples	Examples
Present	Active	501	The media server on which this job is running .
	Passive	117	Specify whether server sets or document sets are being restored.
	Questions	2	Are Backup Exec system services running ?
Modal		22	The cluster service should not be running .
Past	Active	9	When a failover occurs, backup jobs that were running are rescheduled.
	Passive	3	The device was being used by another application (such as a Windows backup utility) when Backup Exec was started.
Infinitive		5	However, my bar code rules don't seem to be working .
Future		2	The Remote Agent will be listening for connections on this predefined port.
TOTAL		661	

Table 2.16: Breakdown of -ing words found introducing the progressive aspect

Referentials

Referential -ing words are those which refer to events or actions (Izquierdo, 2006). As Izquierdo stated, the referential -ing words share the same contexts and syntactic functions as nouns, appearing as subjects, direct objects, verb complements, etc. The subcategories of referentials, therefore, were based on these syntactic functions (see Table 2.17). The -ing words classified as nouns are gerundial nouns. Note that no search was performed to detect gerundial nouns because, as nouns, their context can vary significantly and are difficult to search automatically (apart from the obvious pattern determiner directly followed by an -ing word). Therefore, the number of cases recorded in the classification is incomplete. The total -ing words with referential function retrieved were 594.

Referentials		
Types	Nº examples	Examples
Nouns	252	Configuring a SUDO setting in the bvAgentlessConfig.ini file for query execution
Catenative verbs	167	Symantec recommends performing redirected restore of corrupt files rather than restoring to the original location.
Prepositional verbs	116	The only way to prevent users in a profile from backing up a specific folder is to uncheck this option.
Comparatives	46	Checking the previous job history is faster than performing a pre-scan.
Phrasal verbs	13	This could happen, for example, when the desktop user logs in using a local or cross-domain account.
TOTAL	594	

Table 2.17: breakdown of referential function -ing words

Titles

Finally, let us report on the category added to Izquierdo's (2006) classification to include titles starting with -ing words. Unlike in fiction, titles have a high occurrence in instruction manuals. The translation of titles starting with -ing words is different from the translation of -ing words in running text and their translation therefore, requires the identification of such category. This might cause difficulties for RBMT systems, as they are generally not able to distinguish between running text and titles. It was therefore considered essential to study the performance of our MT system when dealing with this particular structure, particularly given its frequency of occurrence (25% of the -ing words in the corpus).

Within IT documents, titles appear in three different positions. Some appear as independent segments (Free) as headings; others appear within quotation marks at the beginning of sentences (BOS), mainly in introductory indexes where the page numbers for each section are given; and finally some appear embedded in sentences within quotation marks (Embedded) (see Table 2.18). Overall 2,603 occurrences were allocated to this category.

Titles				
Pattern	Position1	Position2	Nº examples	Examples
-ing	Free		1,255	Pausing and resuming devices
	Embedded	Embedded	620	Configure thresholds to recover jobs (" Setting thresholds to recover jobs" on page 490)
		Beginning of sentence	530	" Excluding dates from a schedule" on page 388
About + -ing	Free		100	About restoring NetWare servers
	Embedded	Beginning of sentence	60	"About redirecting Exchange restore data" on page 1260
		Embedded	38	Review the information in "About selecting individual mailboxes for backup" on page 1241.
TOTAL			2,603	

Table 2.18: Breakdown of -ing words at the beginning of titles

2.2 HUMAN EVALUATION

2.2.1 EVALUATION

With -ing words functionally classified, the evaluation was designed. This section reviews the attributes and types of evaluation available and justifies the ones selected for this research to ensure maximum internal and external validity. It also provides a detailed description of the evaluation set-up to enable replication.

2.2.1.1 HUMAN VS. AUTOMATIC EVALUATION

One of the first decisions to make was whether the evaluation would be performed by human judges or automatic metrics. As discussed in Chapter 1 section 1.4.1, traditionally, humans have been the judges of translations. Their competence, particularly of those who have followed additional training is clear. In addition, end-users of translations are also humans.

Nevertheless, as we mentioned, human evaluation is criticised for being time-consuming, expensive, subjective and even inconsistent. Automatic metrics overcome these weaknesses as they constitute a fast, low-cost, objective and consistent option. Yet, one must remember that string-based metrics still require a degree of investment because reference translations – written by humans – are necessary. These

must be commissioned or obtained through existing TMs, for instance. It is precisely by comparing the MT output against these references that the scores are calculated. Moreover, the literature recommends the use of multiple references, up to 3 (Papineni et al. 2002a; Turian et al. 2003) increasing the required investment. No one translation exists for each source text. Variations in grammatical structures and lexicon can result in equally good translations. Therefore, in order to account for legitimate variation in word choice and order, metrics require references which display this variation (Thompson, 1991; Papineni et al. 2002a).

The main challenge of automatic metrics, however, is the coding of concepts of text quality. How can a subjective concept such as fluency be defined in computational terms? As described in Chapter 1 (section 1.4.2) and Chapter 2 (section 2.3.1), researchers have explored different approaches and tested their accuracy by comparing the automatic scores against human judgements. As we said, however, the degree of correlation is not agreed upon. An additional drawback of the current string-based metrics for our research is that the metrics are optimised to score at a text level or at most, sentence level. Their usefulness at subsentential-level has only recently been studied (see section 2.2.1.2).

In view of the problems associated with both human and automatic metrics, we opted for a human evaluation, which would allow us to measure quality attributes and identify the -ing words incorrectly handled by the RBMT system. However, we decided that we would then complement the evaluation by testing whether the most common automatic metrics can indicate the quality of -ing words. As Banerjee and Lavie put it, *“While a single one-dimensional numeric metric cannot hope to fully capture all aspects of MT evaluation, such metrics are still of great value and utility”* (2005: 65-66). Should a correlation exist, we would provide empirical data to support the adequacy of automatic metrics to evaluate linguistic features.

2.2.1.2 TYPE OF EVALUATION

Several human MT evaluation types exist. These vary on the length of the evaluated unit - sentence or constituent - and the type of answer sought - ranking or attribute evaluation. In this section, we describe these types and discuss their advantages and disadvantages to justify our final decision.

ATTRIBUTE ASSESSMENT AT SENTENCE LEVEL

The most widely used methodology to perform human evaluation is the attribute assessment at sentence level (Shubert et al. 1995; Spyridakis et al. 1997; Akiba et al. 2001; Coughlin, 2003; Liu and Gildea, 2005). This is based on defining different attributes such as fluency and adequacy, and asking evaluators to assess candidate translations (sentences or longer texts) on a scale. This type of evaluation provides information about the quality of the sentence or text in general. However, our aim is to identify the particular quality of -ing words, and therefore, this type of evaluation was not considered suitable.

RANKING AT SENTENCE LEVEL

This is a simple evaluation type where evaluators are asked to order different translations of the same source (sentences or longer texts) from best to worst. It is mostly used for MT system comparisons (Turian et al. 2003; Finch et al. 2004). It simplifies the cognitive task significantly as the evaluators do not have to decide on the absolute quality of each candidate (Callison-Burch et al. 2007), resulting in a quicker process. We could evaluate -ing words by asking evaluators to choose between the MT output and a reference translation. However, in the cases where the MT output was ranked as second, we would not know whether this was because the MT output was badly handled or because of stylistic preferences on the part of evaluators.

CONSTITUENT-BASED EVALUATION

Callison-Burch et al. conducted a pilot study of a new evaluation type they called *constituent-based evaluation* (2007). They parsed the source texts and selected constituents for evaluation. The constituent was highlighted in both the source and the translation to facilitate identification and the judges were asked to rank sentences by focusing on that specific part. The evaluation proved that inter- and intra-annotator agreement was higher than that of sentence-level ranking and attribute evaluation tasks. In addition, the speed of the task was reduced from 26 seconds per sentence for attribute assessment and 20 seconds for sentence ranking to 11 seconds for the constituent-based evaluation. Apart from the gain in agreement and speed, this type of evaluation provides the opportunity to target the study and to focus on the constituent of interest instead of having to extract the relevant information from sentence level

scores. Given the reasons outlined above, our judgement was that the constituent-based approach was the evaluation type which best suited our study and, therefore, we adopted it.

The constituent-based evaluation presented by Callison-Burch et al. (ibid) and the evaluation we aimed at performing in this research shared the objective of evaluating subsentential units. This restriction increases the informativeness regarding the errors in the translation. Yet, it differed in the focus. Whereas Callison-Burch et al. were interested in constituents, our objective was a linguistic feature, hence the name featured-based evaluation.²⁵ Despite this, common challenges in the evaluation of subsentential units emerged: the delimitation of the translation unit. In order to address this, Callison-Burch et al. (ibid) defined their constituents with the help of parsed trees. They selected different nodes of the trees and identified the corresponding translation using word alignment techniques. They did not focus on a particular constituent, but rather established conditions for a tree node to be eligible: (1) it could not be a whole sentence; (2) it must consist of 3 to 15 words; and (3) it must have a consistent word alignment in the translation. Our research could not follow the same criteria, as we aimed at studying the -ing words in context but not the context itself. Our approach to delimiting what comprised the translation of the -ing word was to highlight the -ing word and machine translate it. The RBMT system transfers the formatting to the target language, and therefore, this correlated with the approach taken by Callison-Burch et al. (2007). However, they report that *“Because the word-alignments were created automatically, and because the phrase extraction is heuristic, the phrases that were selected may not exactly correspond to the translations of the selected source phrase.”* (ibid: 6). Following their example, evaluators were warned about this possibility.

Highlighting the corresponding translation, however, is not enough to ensure that the evaluators will concentrate only on a linguistic feature. In order to alleviate the difficulty of evaluating the translation of -ing words only and to obtain greater agreement, comprehensive guidelines were written for evaluators (see Appendix B).

²⁵ In MT evaluation the term feature is often used to refer to attributes such as intelligibility or accuracy (Hutchins, 1997). However, this term is also used to mean "grammatical feature", where it refers to a specific linguistic structure. The object of our evaluation being a specific grammatical feature, i.e. -ing words, we name our evaluation *feature-based*.

The guidelines focused on clarifying the type of error that could be attributed to the presence of an -ing word.

Frey et al. (1991) suggest that in order to avoid researcher unintentional expectancy effect, minimum information of the aim of the study should be provided. We concluded that due to the complexity of the evaluation it was preferable to give clear guidelines of what was required and even offer the evaluators the possibility of asking questions. The questions provided us with information on the evaluator's understanding of the evaluation task and they proved very constructive for the analysis of the results. Yet, this communication was established through an independent contact point between the researcher and the evaluators to minimise researcher personal attribute effect. It should be noted that the evaluators were not in contact with one another and therefore could not influence each other's opinions.

An additional issue for this type of evaluation was the overall quality of the sentence. It was thought that a low overall quality might unconsciously influence the judgement of the evaluators towards the highlighted feature. In order to minimise this possibility, two measures were taken. Firstly, efforts were made to improve sentence quality. Project-specific dictionaries were coded following the Symantec in-house procedure. In total, 103 terms for French, 17 for German, 65 for Japanese and 19 for Spanish were encoded.²⁶ Also, it was decided that a post-edited version of the raw MT output would be provided as a guide to what could be considered "correct". Professional in-house post-editors for Spanish, French and German and an external post-editor for Japanese carried out rapid post-editing, that is, changes to respect the target language syntax and lexicon, and for elements that hinder comprehension (Loffler-Laurian, 1996), with special attention paid to -ing words. This would help evaluators get a picture of what could be accepted from the MT system. If evaluators had any doubts about the accuracy of the post-edited version, they were told that they could disagree with it.

²⁶ The low number of terms to be encoded is due to most recurrent domain-specific terms already being encoded in the Symantec user dictionaries. In addition, the documentation for one of the products had already been machine translated into Spanish and German and therefore most of the terminology was already encoded.

Evaluators were asked to judge whether the translation of -ing words was “correct” or “incorrect” based on the following question: Is the machine translation of the ing word grammatical and accurate? (see section 2.2.1.4.1 for a discussion on attributes).

2.2.1.3 EVALUATORS

Evaluators were native language professional translators. We decided not to involve in-house Symantec linguists in the evaluation because they might be biased by their continual involvement in the translation, MT, and development of in-house guidelines and standards. External translators provided by a vendor were hired instead. These translators might have been involved in a Symantec project before but their work was not limited to Symantec projects. Therefore, they had an overview of the expected text quality in different domains, including IT, and were in a good position to act as judges. Additionally, evaluators were asked to complete a questionnaire regarding their linguistic background, professional experience and opinion on machine translation (see Appendix C). All evaluators were full-time translators except 2 Japanese evaluators who, despite doing translation, mainly focused on reviewing. The translation volume of each evaluator varied, ranging from 100 thousand to 5.5 million words, with 7 out of the 16 total evaluators having an experience of over 10 years in translation. Note that 2 French evaluators did not specify the amount of words translated. In addition, we confirmed that all evaluators had taken courses on their native language grammar either at second (5 evaluators) or third level (11 evaluators) institutions, which made them suitable for answering grammar-related questions. Professional experience on post-editing was not shared across the evaluators. 10 evaluators claimed to have performed machine translation post-editing whereas 6 stated that they had never been exposed to any post-editing task. Among the evaluators who had some experience on post-editing, we observed that the amount of post-edited words was quite low, with an average of 10-30 thousand words. We noted that those evaluators who had never carried out any post-editing were one German evaluator and one Spanish evaluator, whereas the remaining four were Japanese. This meant that whereas three or four evaluators had some experience with post-editing for German, French and Spanish, none of the Japanese evaluators had any. Finally, we asked whether they liked working with machine translation. They were presented with a 4 point scale (1 - not at all, 2 - moderate, 3 - somewhat, 4 - very much). 8 evaluators answered "moderate" and 4

"somewhat", whereas only 3 answered "very much" and 1 "not at all". This information ensures the validity and reliability of the evaluators, overriding the possibility of bias due to strong negative or positive attitudes towards machine translation.

A hotly-debated question is what an adequate number of evaluators per language is. A large number is recommendable for generalisability reasons and to limit the impact of subjectivity. However, studies seem to be divided into the use of a high number of student (Shubert et al. 1995; Spyridakis et al. 1997) or non-professional volunteer evaluators (Callison-Burch, 2007; NIST Open Machine Translation Evaluation, 2008; 2009; Zaidan and Callison-Burch, 2009), and a low number of professionals ranging from two to four (Aiken and Wong 2006; Lagarda, et al. 2009). Threats to the validity of the results are addressed by incrementing evaluator numbers for the former, whereas the latter avail of professionals with the view that their responses will be more objective and reliable. Due to the complexity of our evaluation, and to ensure reliable responses, we opted for expert evaluators. In order to decide on the exact number, a balance between the number of evaluators, the language-pair span and the cost had to be sought. It was one of the aims of this research to explore translation improvement techniques for both pre- and post-processing stages. Particularly for CL, it is essential that the proposed rules are beneficial – or at least neutral – for the target languages into which texts are translated. Therefore, it was necessary to perform the evaluation for a number of language pairs. After considering our requirements and the budget, we decided to hire 4 professional translators for 4 target languages, that is, 16 in total. The number of evaluators fell into the accepted range for professional judges and the span of target languages used would provide us with data to make reliable suggestions for pre-processing techniques.

As an additional measure to monitor the reliability of the results obtained from the evaluators, the inter-rater agreement was calculated through the Kappa statistic (Fleiss, 1971; Carletta, 1996; Callison-Burch, et al. 2007). The Kappa coefficient is a calculation based on “*the difference between how much agreement is actually present (‘observed’ agreement) compared to how much agreement would be expected by chance alone (‘expected’ agreement)*” (Viera & Garrett, 2005):

$$K = \frac{P_{obs} - P_{exp}}{1 - P_{exp}}$$

Figure 2.2: Equation for the Kappa coefficient, where P_{obs} is the observed agreement and P_{exp} is the expected agreement.

Kappa scores can vary from -1 to 1, where 1 is perfect agreement, 0 is agreement due to chance, and -1 is perfect disagreement. French, German and Spanish showed good agreement with 0.702, 0.630 and 0.641 respectively. Japanese showed slightly lower agreement, although still moderate, with 0.503. These results proved the validity of the evaluation set-up and indicated that the evaluators had a similar understanding of the task.

2.2.1.4 EVALUATION QUESTIONS

Two issues were to be considered for designing the evaluation questions. First, the evaluation attributes had to be selected and next the response format chosen.

EVALUATION ATTRIBUTES

The most popular attributes for MT evaluation are fluency and adequacy (LDC, 2003; Popescu-Belis, 2003; Callison-Burch et al. 2006; Callison-Burch et al. 2007; Hamon et al. 2007). Fluency refers to “[t]he extent to which a sentence reads naturally” (FEMTI). It is also called readability, intelligibility and clarity (ibid). But what are the variables that determine how well a sentence reads? The metrics proposed by FEMTI vary from the rating of sentences in scales to cloze tests, comprehension tests based on questionnaires or reading time measurements. The rating of sentences leaves the interpretation of the meaning of fluency up to the evaluators, as it does not further explain the variables to be considered. Cloze tests and comprehension tests emphasise variables related to understanding, which, in turn, it could be argued, depends not only on the intrinsic quality of the sentence but also on the readers’ capacity. The intrinsic qualities that help understanding could be grammar, domain-specific style, cohesion and coherence, for instance. Reading time might depend on the expectations of the readers regarding grammar and style. Grammaticality is defined by FEMTI as the “degree to which the output respects the reference grammatical rules of the target language” (FEMTI). Style, in turn, is “a subjective evaluation of the correctness of the style of each sentence” (van Slype in FEMTI) “also commonly referred to as “register”

and includes degree of formality, forcefulness and bias as exhibited through both lexical and morpho-syntactic choices” (FEMTI). However, FEMTI notes, style and fluency are *distinct*. This would mean that the degree of naturalness of a sentence is not measured by the domain-specific terminology and phraseology used, which might be counter-intuitive for an evaluator. To sum up we could say that whereas fluency is a well-established attribute in translation evaluations, the definition of the concept is not yet clear.

Given the conceptual inconsistency behind the term fluency and because we deal with an MT system based on grammatical rules, we opted for the evaluation of grammaticality. We also believe that whereas judging style or comprehension can be subjective and idiosyncratic, grammaticality is a more objective attribute. As a result, its measurement requires less cognitive effort and should improve inter- and intra-evaluator agreement. Grammaticality would also be a more workable attribute to be judged when evaluating the translation of a grammatical feature as opposed to a unit with a complete meaning. Therefore, we present the evaluators with specific contexts and ask them whether the translations of -ing words were grammatical in those particular contexts.

Adequacy, also called accuracy, fidelity, correctness or precision (FEMTI) refers to the “subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation” (Van Slype’s Critical Report). We believe that a grammatical rendition of a source text is of little use if the meaning conveyed is different. We are aware of the fact that a feature might not always provide enough information to judge whether the meaning is being conveyed or not. However, due to the importance of this attribute it was incorporated into the question to pinpoint the obvious cases where meaning was lost or was different.

The question asked of the evaluators was the following: Is the machine translation of the -ing word grammatical and accurate?

RESPONSE FORMAT: SCALES VS. BOOLEAN QUESTIONS

The use of scales in MT evaluation is widespread (LDC, 2003; Callison-Burch et al. 2007). They provide the possibility to rate across a range of grading parameters which makes it possible to calculate the average perception of quality. However, as

Greenhalgh (1997) claims, *“the numerical result would be uninterpretable unless we knew that the difference between ‘not at all’ and ‘a bit’ was exactly the same as the difference between ‘a bit’ and ‘a lot’”*. In addition, how do judges decide what makes a translation improve from a 3 to a 4 or a 5 score?

Given the subjectivity of scales, a Boolean response was considered for two main reasons. First, as Cavar et al. claim “[a]s is well known from experimental psychology and psycholinguistics, simple binary decision tasks (e.g. yes/no questions), for example, are answered much faster, and more reliably across items and across subjects by the evaluators than decision tasks that provide a decision scale” (2000: 1). Secondly, our main interest was to obtain a list of ungrammatical and inaccurate renderings of subcategories of -ing words in order to explore techniques to improve their translation. Therefore, the use of scales was not considered necessary. Instead we offered a simple “yes/no” response choice to evaluators.

2.2.1.5 LANGUAGE SELECTION AND MT PROCESS

Although we are aware that for post-processing techniques each individual target language will have to be analysed, for pre-processing efforts, where common issues are addressed, a wide coverage of the different languages is necessary. For this reason, the languages selected for the study were French, German, Japanese and Spanish. These languages covered the range of Romanance, Germanic and Altaic families and are some of the most common target languages for Symantec.

Once the user dictionaries were updated and the domain-specific dictionaries loaded, the standard settings for IT domain (independent translation of quoted segments, polite forms, infinitive instead of imperative forms) were selected for the MT stage. The main feature to mention here was the choice of the polite treatment for second person singular pronouns and verbs, particularly for the imperative forms.

2.2.1.6 SAMPLING

The choice of performing a human evaluation determined the evaluation set-up because on the one hand, the cognitive effort required from the evaluators would have to be taken into account and, on the other, the expense involved in hiring their services would also be an aspect to contemplate, as already mentioned. Therefore, the balance between three factors had to be sought: cognitive effort, generalisability of results and

budget. In other words, we had to select a set of sentences for evaluation which would provide us with enough results to draw solid conclusions while being feasible for the evaluators to answer them in a span of time that would fit into our assigned budget. The challenge that emerged at this point was that of finding a sampling method which would ensure the external validity of the evaluation and at the same time allow us to perform the evaluation within the specified conditions.

Different sampling methods have been applied to linguistic studies. These range from random sampling techniques, which best reflect the characteristics of a corpus and where all subjects of the population have equal chances of being selected, to non-random techniques, usually used when the latter are not possible due to budgetary constraints or the impossibility of finding suitable population subjects (Frey et al. 1991). The purest random sampling technique is called simple random sampling (SRS). It is based on assigning a consecutive number to each of the subjects in the population and selecting them randomly until the required size is obtained (ibid). These techniques are often used to representatively reduce the total population to perform a more or less detailed analysis of the relevant feature(s) and classification (Kretzschmar, et al 2004; Izquierdo, 2006). Cochran (1963) proposed a model (see Figure 2.5) for obtaining a representative sample of proportions for SRS. This model requires specifying a confidence level, that is, how sure you are about the results being correct, and a confidence interval, that is, the error-margin expected from the result. It allows for conclusions in the line of: *we are 95% sure that -ing words are correct 56% ±5 of the time*. It makes two assumptions: firstly, it can only be used for Boolean questions and secondly, the sampling design with which it must be used is SRS (ibid). This model also accounts for the degree of variability of the sampled data, that is, whether the units to be evaluated are similar or are thought to be fairly heterogeneous (see Figure 2.3).

$$SS = \frac{Z^2 p (1 - p)}{C^2}$$

Figure 2.3: Formula to calculate sample size (SS), where Z is the confidence level value (e.g. 1.96 for 95% confidence level), p is the degree of variability (0.5 used, maximum variability), and C is the confidence interval, expressed as a decimal (e.g. 0.04 = ±4).

The use of this formula to select a representative sample of the sentences containing -ing words was considered. 385 sentences would have sufficed to draw a

conclusion on the whole corpus with a 95% confidence level and ± 5 interval, assuming maximum variability of the sampled data.²⁷ This would have been acceptable for our limitations on budget and would probably avoid excessive cognitive effort. However, it was not thought of as satisfactory because the possibility of drawing conclusions on more specific subcategories of -ing words was minimal and, from the pilot project, we learnt that some subcategories were usually considered to be translated adequately, whereas others were systematically inadequate. In addition, two more technical details failed to prove appropriate for our study. Firstly, SRS assumes that the total population is known, which is not true in our case. We are using a corpus (already a sample) and using 80% of it. Secondly, the variability value to be applied highly depends on whether we expect the difference between adequate or inadequate quality to be reported by the evaluators. It was suggested that due to the lack of expectations and the impact of this value in the sample size, Cochran's model and in general SRS was not ideal for our study. Therefore, the suitability of a stratified sampling method was considered.²⁸

Several corpus studies have used the stratified sampling approach (Dagan et al. 1993; Hochberg et al. 1998; Kretzschmar et al. 2004; Areta et al. 2007). This approach is usually reported to allow the researcher to select and focus on the relevant characteristics of the subjects of study. Subjects are selected from subgroups (strata) of the total population. These subgroups are based on observed characteristics of real occurrences and have been created by classifying the subjects according to non-overlapping characteristics relevant for the study. In fact, Hochberd et al. claim that *“one can generally reduce sampling uncertainty by combining results from several homogeneous strata, rather than doing an overall sample from a heterogeneous population”* (ibid: 1998: 1). This approach was thought to be more convenient for this research as the characteristics for the strata were easily identifiable and the results would allow for detailed conclusions on specific -ing word subcategories. Once the strata were ready, the sample was obtained by selecting the subjects randomly and in proportion in each group. This allowed us to identify the most problematic -ing word

²⁷ Note that the confidence interval greatly changed the number of examples required. A confidence interval of ± 1 would require 9,604 examples, whereas ± 2 ± 3 ± 4 ± 5 would require 2,401, 1,111, 600 and 285 examples respectively.

²⁸ In consultation with Dr Lynn Killen and Mr. Gary Keogh from DCU School of Computing.

subcategories for RBMT systems in the IT-domain, which must take into account both incorrect translations and occurrence rates. Because subjects are selected proportionately, larger subcategories would have proportionately more sentences included in the sample.

Stratified sampling has its own deficiencies which should be pointed out here. In contrast to Cochran's model, when using stratified sampling the confidence levels cannot be predetermined, which adds a degree of uncertainty at the time of carrying out an evaluation. Of course, once the results are known the confidence levels can be calculated. This leads to different strata having different confidence levels and therefore comparisons require a more complex treatment.

In order to proceed with the stratified sampling, the number of subjects to be extracted had to be decided. From the pilot project we learnt that evaluators took 20-24 hours to evaluate 1,857 sentences. They described the task as demanding but agreed that it was feasible with enough time to go through all sentences once. Our statistics expert advised us that 1,800 instances out of 8,363 was a considerably high ratio so this number was set as the required sample size.

As McEnery et al. (2006) point out, when using stratified sampling, one must ensure that the number of subjects selected from each stratum are proportional to their occurrence rates in order for the sample to be considered representative. Following the example of Kretzschmar et al. (2004) a systematic sampling technique was used to select the subjects. In systematic sampling, subjects are selected every n th from each stratum until the sample quota is reached. Note that it is considered a random procedure because it starts at a random point (Kochanski, 2006) which further increases the external validity of the method.

A risk associated with using systematic sampling is periodicity. This is a phenomenon that emerges if, for some unexpected reason, all the n th subjects share a particular feature which can bias the results obtained from the sample. Although the strata were grouped without any ordering feature in mind, to avoid all potential bias, the sentences in each stratum were ordered by sentence length. This would rule out any possible periodicity and also encourage different sentence lengths to be included in the sample.

Assuming that our strata contain all -ing word subcategories in the corpus with the respective characteristics, every 5th subject was selected (1,664 subjects) and in a second round every 60th until the sampling quota of 1,800 -ing words was reached (see Table 2.19 for the final sample and Appendix D for list of abbreviations).

-ING Word Sample					
TITLES – 557					
Title_ING_i ndp	Title_ING_Q M2	Title_ING_Q M1	Title_ABING _indp	Title_ABING_Q M2	Title_ABING _QM1
270	132	113	22	8	12
CHARACTERISERS - 536					
Char_PR	Char_POrr	Char_POnn	Char_POadj		
401	84	49	2		
ADVERBIALS - 414					
Adv_M_by	Adv_M_with out	Adv_M_0	Adv_M_with	Adv_CT_instead of	Adv_T_before
111	19	25	X	3	38
Adv_T_afte r	Adv_T_while	Adv_T_when	Adv_T_betwe en	Adv_T_upon	Adv_T_on
30	14	67	1	X	2
Adv_T_in	Adv_T_durin g	Adv_T_prior	Adv_T_along with	Adv_T_inthemi ddleof	Adv_T_from
X	X	X	1	1	1
Adv_T_thro ugh	Adv_T_0	Adv_CC_besi des	Adv_P_where	Adv_PU_for	Adv_PU_0
1		X	X	95	X
Adv_CD_if	Adv_C_becau se				
4					
PROGRESSIVES - 141					
Prog_PRact	Prog_PRpas	Prog_PRq	Prog_PSact	Prog_PSpas	Prog_fut
107	26	X	1	1	X
Prog_mod	Prog_inf				
5	1				
REFERENTIALS - 134					
Ref_nn	Ref_comp	Ref_prepV	Ref_phrV	Ref_cat	
62	10	24	2	36	

Table 2.19: Sample of the -ing words obtained by applying a stratified systematic sampling method.

2.2.2 ANALYSING FRENCH AND SPANISH

This section sets out the methodology used to examine the structures generated by the RBMT system for each subcategory in the English-French and English-Spanish language pairs and their correctness according to evaluators. The analysis will be restricted to the general trends within each subcategory because of space constraints.

A few aspects need to be considered with regard to the analysis methodology followed. Firstly, the grouping of scores must be presented. From the binary question and 4 evaluators, each -ing word translation could get 4 to 0 points. Following our binary approach to results, we thought appropriate to consider translations scoring 3 or 4 points, i.e. 3 or 4 evaluators rated them as correct, as correct and translations scoring 0 or 1 point, i.e. none or 1 evaluator rated them as correct, as incorrect. -ing word translations scoring 2 points lie mid-way on the scale and could only be interpreted as being inconclusive. We acknowledge the somewhat arbitrariness of this approach but we believe it is acceptable to report the majority vote. For the more detailed analysis, however, we differentiate between each scoring by considering translations with 3 and 1 votes correct and incorrect, respectively, but with a lower confidence level.

Secondly, the approach to the linguistic reasoning behind the results must be described. Whereas it was plain to see why the evaluators considered an -ing word translation correct, the exact reason why it was considered incorrect was not always obvious. The evaluation guidelines stated that the translation should be considered correct if it was grammatical and accurate. Therefore, the translation was incorrect if it was ungrammatical and/or inaccurate.

Three sources of difficulty in pinpointing the specific reason for the “incorrect” score arise from this definition. To start, grammars are not always exhaustive and do not give every possible usage combination for a particular feature. As a result, grammar and usage seem to intertwine to the extent that it could be argued that usage “fine-tunes” grammar. The possibility exists that structures seem ungrammatical just because speakers do not use them and they sound alien to the language. The boundaries between grammatical and unnatural might change from evaluator to evaluator.

Moreover, as was mentioned before, delimiting the exact word(s) pertaining to the translation of the -ing word was complex. In order to control the evaluated words, the evaluation guidelines included explanations and examples of what should and what should not be penalised. Clearly, it was not feasible to provide an exhaustive account of dubious contexts.

Finally, despite our goal being the evaluation of a specific grammatical feature, it was necessary to introduce a concept that would consider both source and target

languages to make sure that the target was the correct translation of the source. Accuracy was chosen to cover this concept. However, accuracy also encompasses the appropriateness of the terminology. Although measures were taken to ensure that domain-specific terminology was used, the possibility of terminological inaccuracies still existed, and might have swayed evaluators towards a negative rating.

2.2.3 ANALYSING JAPANESE AND GERMAN

The analysis of the evaluation results for German and Japanese followed a different methodology from the one used for French and Spanish as the author is not proficient in the former languages. We could of course have limited our study to the two target languages the researcher is proficient in. However, breadth of coverage was important for the industrial sponsor. Therefore, we included third party analysis for German and Japanese, but tried to control this activity closely. For the analysis of French and Spanish a description of the translations of -ing words was carried out. All evaluated instances were considered and the different translations generated for each -ing word subcategory discussed. This resulted in a detailed description of the MT system's behaviour when dealing with different -ing word subcategories. This description was contrasted against the grammaticality and accuracy judgements of evaluators. Finally, the most recurrent issues and problematic subcategories were singled out for improvement.

However, for German and Japanese, we restricted the analysis to identifying the main errors for these two languages, as well as the most affected subcategories. This would allow us to create a list of recurrent errors and difficult subcategories which would then be used to compare against the results for Spanish and French. With this information, we could decide what type of action could be taken to solve them.

A mother tongue linguist per target language was asked to analyse all instances of German and Japanese classified as incorrect by the evaluators.²⁹ In order to do so, we created a list with the most recurrent errors in Spanish and French, which was provided in the analysis spreadsheet together with the -ing words evaluated as incorrect. These

²⁹ The Japanese linguist was a PhD student in translation studies, working on the area of machine translation and post-editing. The German linguist was a Symantec employee, who coordinates the efforts carried out to improve and maintain the quality for the German language in the human and machine translation workflows at the localisation department.

errors, as well as a brief summary of the evaluation questions, were described in the Analysis guidelines (see Appendix E for the complete Analysis guidelines). The linguists could choose between the listed errors (e.g. preposition error, loss of progressive aspect, etc.) while supporting their judgements with information on the specific words/characters involved (specifying the incorrect preposition and providing the correct one, for instance); and additionally, they could describe new errors which appeared in their target languages and were not covered in the list.

2.3 AUTOMATIC EVALUATION

The requirements for an ideal automatic metric were described by Banerjee and Lavie (2005). According to them, such a metric should, above all, correlate highly with human evaluations. Additionally, the metric should also be able to report minor differences in quality both between different systems and updates of the same system. For comparison purposes, this would have to be done in a consistent manner. Equally, it would have to be possible to assume a similar performance for systems with similar scores. Finally, they mention the importance for the metric to be general in the sense that it can be used for different MT tasks, domains and scenarios.

As the same authors admit, meeting the above-mentioned requirements is very difficult and the attempts carried out so far have not managed to satisfy them (ibid: 66). Focusing on our experiment, it is not our aim to create a new automatic metric, but rather to review the most widely used ones to examine whether they would be useful in the scenario set in this project. In detail, we are looking for a metric which may correlate with the human judgements of an aggregate of accuracy and grammaticality at a subsentential level, where scores must be obtained for texts for which English is the source language and French, German, Japanese and Spanish are the target languages.

2.3.1 AUTOMATIC METRICS

BLEU

The first attempts to automate MT quality evaluation tried measuring the number of keystrokes required to convert MT output into a reference translation (Frederking and

Nirenburg, 1994). But it was not until IBM presented their automatic metric BLEU (Papineni et al. 2002) that extensive work started to appear in the field.

The authors acknowledged the capacity of humans to perform MT system quality evaluations and their usefulness. However, they spotted what they thought hindered a more rapid progress of MT development: human evaluation was expensive and slow for developers to test the effects of "daily" changes to their systems. Therefore, they borrowed the baseline of the Word Error Rate (Olive, 2005) metric from the speech recognition community and proposed BLEU (bilingual evaluation understudy).

This metric is based on precision between MT output and several reference translations. It calculates the number of different n -grams (usually up to 4) in the MT output which match with the n -grams present in the reference translations and divides this by the total number of n -grams in the MT output. It uses what the authors call a "modified n -gram precision" to avoid a particular n -gram in the MT output being assigned to more than one n -gram in the reference translations. The number of times a particular n -gram appears in a single reference is counted. Then the total count of this particular n -gram in the MT output is counted and divided by the number obtained from the reference translation. According to the authors, this modification of precision makes it possible to account for both "adequacy" – as it accounts for the words in the reference present in the MT output – and "fluency" – as higher n -gram matches account for word-order measurement. As for the question about how the scores for different levels of n should be combined, the authors propose the average logarithm with uniform weights, equivalent to the geometric average, that is, the average scores for each n -gram level are multiplied. However, this points to one of the weaknesses of the metric. In the cases where the score for a particular n is zero, the metric does not consider the values obtained for the other n and reports zero.

The description of the metric so far accounts for precision but nothing has been said about recall, that is, measuring the amount of words in the reference translation also present in the MT output. However, the authors admit that computing recall when dealing with several references is difficult. Usually, a high recall means that many of the words in the reference translation are included in the MT output. However, MT output with additional words which happen to be included in the different reference translations would also score highly, whereas the quality could be low. Therefore, in

order to control for MT sentences that are too long, they introduce a brevity penalty. This is done at a text level to allow for certain freedom at sentence level.

The experiment presented by the authors showed very high correlations with a human evaluation of readability and fluency. Specifically, the correlations were 0.99 for monolingual judges and 0.96 for bilingual judges (Papineni et al. 2002). The difference might be due to the fact that bilingual judges might have penalised issues regarding accuracy of the information transferred, that is, recall (although the notions of readability and fluency do not address this directly). Figure 2.4 outlines the equation for BLEU.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$BLEU = BP * \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

Figure 2.4: Equation for BLEU, where the modified n-gram precision p_n is calculated for n-grams up to number N and using positive weights w_n and it is multiplied by the exponential brevity penalty factor BP where c is the length of the MT output and r is the length of the reference text translation.

NIST

The Defense Advanced Research Projects Agency (DARPA) commissioned the National Institute of Standards and Technology (NIST) to consider the newly created BLEU metric. It was to be used in the DARPA-funded Translingual Information Detection, Extraction and Summarization (TIDES) programme. In doing so, NIST examined the IBM metric and improved it, obtaining better correlations with human judgements of adequacy and fluency (NIST report, 2002) by the metric since called NIST.

Two main modifications were applied. First, the arithmetic average was used to combine the scores for different levels of n-grams. This removes has the effect that the scores for different levels of n-grams not to depend proportionally on one another. Secondly, less frequently occurring n-grams were given more weight than the frequently occurring ones. Assuming that less frequently occurring n-grams add more information, adequacy should benefit from this decision.

$$NIST = \sum_{n=1}^N \left\{ \frac{\sum_{\substack{\text{all } w_1 \dots w_n \\ \text{that co-occur}}} Info(w_1 \dots w_n)}{\sum_{\substack{\text{all } w_1 \dots w_n \\ \text{in sys output}}} (1)} \right\} * \exp \left\{ \beta \log^2 \left[\min \left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}$$

Figure 2.5: Equation for NIST, where β is the brevity penalty factor = 0.5 when the number of words in the system output is 2/3 of the average number of words in the reference translation, N is 5 and L_{ref} is the average length of the reference translations and L_{sys} is the length of the MT translation.

It is interesting to note that during the examination of the metric, NIST reported, among other factors, the effect of the number of references and the segment size on the metrics' scores. Their experiments showed that using more than 1 reference translation did not improve the correlation with human judgements significantly. On the segment size, they reckoned that obtaining co-occurring n-grams was easier for shorter segments, and therefore the score would be higher, and vice versa. Also, they suggested that aligning MT output to the corresponding translation was more difficult and even unnatural as the segment size was reduced and concluded that segments should never be shorter than one sentence.

GTM

In the following year, Turian et al. (2003) pointed out one of the limitations of BLEU and NIST. Although they could be used to rank different systems or different versions of the same system, the scores did not identify specific errors to guide developers in improving the systems. Therefore, they claimed that MT could be evaluated using the standard precision and recall and their composite F-measure, which, thanks to its intuitive graphical representation, did offer insights into problematic sequences. They called their proposal General Text Matcher (GTM).

Precision measures the number of words generated by the MT system that match with words in the reference translation out of the total number of words generated by the MT system. Recall measures the number of words generated by the MT system that match with words in the reference translation out of the total number of words in the reference translation. In order to compute them, the authors borrow the concept of “*maximum matching*” (Cormen et al. 2001: 1051) whereby words in common between the MT output and the reference translation are counted without allowing for a single word to be matched more than once. However, these measures score similarly

regardless of the ordering of the words. Thus, the authors reward matching adjacent words by calculating the square root of the possible adjacent sequences to choose the longest sequence (run) (see Turian et al. 2003 for details).

$$Fmeasure = \frac{2PR}{P + R}$$

$$Fmeasure = \frac{2 \left(\frac{MMS(CT, RT)}{|CT|} \right) \left(\frac{MMS(CT, RT)}{|RT|} \right)}{P + R}$$

$$Fmeasure = \frac{2 \left(\frac{\sqrt{\sum_{r \in M} length(r)^2(CT, RT)}}{|CT|} \right) \left(\frac{\sqrt{\sum_{r \in M} length(r)^2(CT, RT)}}{|RT|} \right)}{P + R}$$

Figure 2.6: Equation for GTM, where the Fmeasure is the composite of P (precision) and R (recall) and where the intersection of the elements CT (candidate text) and RT (length of the reference text) are computed using an extended form of the MMS (maximum matching size) which rewards groups of longer adjacent matching words r (run).

METEOR

Building on the precision and recall measurement approach, Banerjee and Lavie (2005) proposed METEOR (Metric for Evaluation of Translation with Explicit Ordering). The authors aimed at addressing four weaknesses identified in BLEU: the lack of recall, the lack of explicit word-to-word mapping between a translation and its reference, the use of higher n-grams as a measure for grammaticality, and the averaging method to obtain the final score.

In a first alignment stage, METEOR maps each unigram in the MT output to the unigrams in the reference translation so as to obtain a one to one, or one to zero mapping. This means that one unigram in a string cannot map to two or more unigrams in another string. In order to establish the mapping, METEOR can use three different modules. The “exact” module maps unigrams based on their surface form. The “porter stem” module maps unigrams that are the same after stemming. Finally, the “WN synonymy” module maps unigrams if they are synonyms of each other according to the WordNet lexical database. This alignment helps measure the degree of precision and

recall between the MT and reference strings. In order to find the most appropriate combination of the two measurements, the authors carried out several experiments which showed that for the combined score of fluency and adequacy by humans, precision correlated less than recall (ibid: 69-71). Therefore, they apply a harmonic-mean (based on van Rijsbergen, 1979), heavily emphasising the value of recall, to obtain the Fmean.

The equation so far does not account for any explicit measurement of well-formedness. The authors disagree on the use of high n-grams to measure this. Instead, they present a new proposal. They assume that the higher the number of adjacent unigrams an MT output has in common with a reference translation, the better well-formed it is. Therefore, they count the minimum amount of unigram “chunks” in which a reference translation must be fragmented to map to the MT output. The longer the adjacent n-grams are, the fewer the chunks are and the better the score is. The penalty for well-formedness is restricted to a maximum of 50% of the total score. From the equation we can also conclude that the metric, as opposed to BLEU, does not penalise short segments, as no geometrical average is used to compute precision and recall and the penalty can never be zero.

$$Score = Fmean * (1 - Penalty)$$

$$Score = \frac{10PR}{R + 9P} * \left(1 - 0.5 \left(\frac{\# chunks}{\# unigrams_matched} \right) \right)$$

Figure 2.7: Equation for METEOR, where the Fmean is the combination of *P* (precision) and *R* (recall) and a maximum Penalty of 0.5 is introduced to account for non-adjacent unigrams.

TER

In 2006, Przybocki et al. (2006) decided to change the evaluation paradigm and returned to an edit-distance approach to MT evaluation for the NIST Machine Translation Evaluation for GALE (Global Autonomous Language Exploitation). As mentioned above, the idea of using some measure of edit distance as a metric for MT quality had been previously explored but was disregarded after the innovative measurements proposed in the following years. However, the advantages of the new approaches and their roles started to be re-considered (Callison-Burch et al. 2006). The

new metrics did not seem to adequately assess the quality of MT system translation, nor quantify the usefulness for an end user (Przybocki et al. 2006).

In an effort to address these issues, Przybocki et al. (2006) adopted the newly developed software by Snover to calculate edit distance between the MT output and a reference translation. Snover et al. (2006) describe TER (Translation Edit Rate) as “*the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length of the references*” (ibid: 225).³⁰ The edits can be insertions, deletions, and substitutions of single words and also shifts of word sequences. The penalties are the same for all edits. Punctuation marks are treated as a word and differences in capitalisation count as an edit.

TER is calculated in two steps. First, the total number of edits is counted by using dynamic programming to search for the best combination between insertions, deletions and substitutions and shifts. Then the minimum distance is calculated with a cost of 1 for each edit. In the case where more than one reference is available, a score is calculated for each of them and the best is used. We find similarities between TER and the maximum matching size (MMS) method used by Turian et al. (2003) for GTM in that neither of them allows a word to be matched more than once and both allow reordering. However, in contrast to MMS, TER does not explicitly favour longer word matches and the cost assigned to phrasal shifts is lower than in MMS.

$$TER = \frac{\# \text{ of edits}}{\text{average \# of reference words}}$$

Figure 2.8: Equation for TER, where the number of edits is divided by the average number of words in the reference.

Due to the dependency of the metrics upon the reference translations, Snover et al. (ibid) explored the results of the TER metric with specifically targeted translation references (HTER). They asked fluent speakers of the target language (monolingual) to create a version of the MT output which preserved “*the meaning of the reference translations*” with the least possible edits (ibid: 227). This approach proved to reduce

³⁰ The GALE community uses the acronym TER as a short form for Translation Error Rate, derived from the Word Error Rate (WER) metric used in the automatic speech recognition community. Snover et al. (2006) do not agree with the implication of the terminology that it refers to a “definitive” MT measure and use TER as a short form for Translation Edit Rate.

the TER score by 33%, obtaining the highest correlation with human judgements in comparison to BLEU or METEOR and their human targeted versions (ibid: 28-31). Yet, the authors acknowledge that HTER would be too time-consuming (a human takes 3-7 minutes per sentence to create the closest version) and expensive to implement in a development cycle.

Edit-distance was established with the aim of measuring PE effort. The measurement of PE effort is a good guide to estimate the quality of the MT output as well as to identify the problematic words. The count of the minimum keystrokes that are necessary to transform the MT output into a good quality translation to account for the PE effort, however, is quite limited. Researchers have shown that PE effort encompasses a tri-dimensional task and three aspects must be accounted for to report it: the technical effort, the temporal effort and the cognitive effort (Krings, 2001; O'Brien, 2006). It is the technical effort, described by Gennrich (1992) as the “*physical operations that must be performed on a machine translation to correct any errors or defects*” that edit-distance measures, leaving temporal and cognitive effort unaccounted for.

OTHER METRICS

Several different automatic metrics were proposed in parallel with the above-mentioned ones (WER, Nießen et al. 2000; PER, Leusch et al. 2003; ROUGE, Lin and Och, 2004; etc.). However, those described in detail here have been the metrics adopted in most MT evaluation campaigns (CESTA 2007, ACL Workshop on Machine Translation 2007, NIST Metrics for Machine Translation Challenge 2008) and have therefore been maintained and updated consistently. It is worth mentioning that all the metrics examined belong to the same strand of automatic metrics, that of the string-based variety. Nevertheless, work is also emerging from other approaches such as dependency-based metrics and machine learning-based approaches. The former try to go beyond string comparison and use information on the syntactic structure of the MT output and the reference translation (see Liu and Gildea, 2005; Giménez and Màrquez, 2007; Owczarzak et al. 2007a, Owczarzak et al. 2007b) or even MT output alone (see C-score, X-score and D-score, Rajman and Hartley, 2001) whereas the latter experiment with linear regression models, for instance (see Russo-Lassner et al. 2005;

Albrecht and Hwa, 2007). The examination of all these metrics and the different strands goes beyond the scope of this project.

2.3.2 CONSIDERATIONS FOR THE EXPERIMENT

We decided to test the capacity of the readily-available, most commonly used string-based metrics to correlate with human judgements of “correct” and “incorrect” -ing words. From the automatic metrics discussed above, we discarded BLEU because of the averaging technique the metric uses. Once the scores for n 1-4 are calculated, they are multiplied. This means that if any n scores 0 because the segment is shorter than this particular number of n , the total BLEU score is 0. The translations for the -ing words in our target languages varied from 1-6 words and, therefore, most of the examples would obtain a 0 score. Also, it is generally agreed that 4 reference translations are required for BLEU to perform well and its best-performing evaluation unit is the text.

We mentioned that the metrics are optimised to calculate the scores of English MT output whereas in our experiment English is the source language. The only exception is METEOR, which is available to calculate scores for French, German and Spanish as well as English. Note that a Japanese model is not available and therefore, we were not be able to use METEOR for Japanese.

We were also aware that due to the short length of the strings, the different possible scores obtained from the evaluation metrics would be limited, compared to using longer segments. Therefore, and in order to check whether the possibility of allocating a wider range of scores would provide a better correlation with human judgements, we decided to include a character-based metric. Edit-distance, also called Levenshtein distance and originally from the field of information theory, was conceived as a character-based metric to calculate the difference between two strings. As we saw, TER is based on this metric, but applied at a word-level instead. We decided that we would add the character-based edit-distance to the automatic metrics to be calculated to examine differences between word-based and character-based metrics.

2.3.2.1 CORRELATIONS

Rather than obtaining a particular automatic score for our data, we were more interested in examining whether correlations with the human judgements existed. This

is calculated with correlation statistics, mainly the Pearson r or Spearman ρ statistic. Both report the associations between two variables but their assumptions are different. The Pearson r correlation assumes that the data in each variable is interval and normally distributed. Spearman, in turn, only assumes that the data is ordinal, that is, it does not place importance on the values but on the ranking of the data. Therefore, it could be said that Spearman r is basically a Pearson's r of the ranks. The weakness of Spearman is that it is affected by ties. As Richards, (1998) states, *"this will be especially likely to happen when your variables have a small number of possible values and you have a large number of cases"*. He goes on to say that the score for Pearson's r and Spearman's ρ would be very close when few ties exist. Therefore, we were in a position where our data was not purely interval and the variables have a small number of possible values. We say that our data is not purely interval because the human evaluation scores are based on quality concepts which have been judged in a binary question where each evaluator assigned a point per answer. We say that the variables have a small number of possible values because the evaluators could only answer "correct" or "incorrect" and due to the low number of words included in each -ing word translation, the scores reported by automatic metrics were also limited. Our statistics expert, however, informed us that given the large amount of data and the robustness of the statistics, these conditions would not influence our results and we could safely report both correlation coefficients.

We looked at how correlations were reported by the MT community. Both statistic tests were found but Pearson seemed to be more frequent (Snover et al. 2006; Barnejee and Lavie 2005; Owczarzak, 2008). Yet, the adoption of one or other correlation metric does not seem to be standardised. For the NIST Open Evaluation 2008 the group started off by calculating not only Pearson and Spearman correlation but also Kendall's tau. Given the lack of a standard in the MT literature, and due to the larger number of reports for Pearson, we decided to use this metric for comparison purposes, although we also report Spearman's ρ and Kendall's tau correlation coefficients.

2.3.3 EXPERIMENT SET-UP

Our aim was to investigate whether the most commonly used automatic metrics could be used to perform the evaluation of subsentential units and whether the results

correlated with human evaluation results. Automatic metrics require a reference translation, be it a human translation or a post-edited version, against which the MT output will be compared (Papineni et al. 2002a; Tyers and Donnelly, 2009). Professional translations tend to be freer whereas post-edited versions tend to follow the MT output closely. As a result, the similarities between MT output and PE versions are greater, which result in higher scores. As mentioned above, we are not interested in the actual scoring, but in the correlations with human judgements. We believed that this would not be influenced by performing the calculations against PE versions and therefore the post-edited version of the 1,800 sentences used during the human evaluation was used.

In order to perform the calculation at the feature level and compare it to the human evaluation, it was necessary to isolate the exact string considered by the human evaluators. In these isolated strings, we tried to include all and only the elements which human evaluators were asked to judge. As described in section 2.2.1.2, by machine translating the source sentences with the -ing words highlighted, the MT system mapped the formatting to the appropriate words in the target text. Similarly, when post-editors were asked to edit the MT version, they were asked to maintain (or improve) this alignment by highlighting or by removing the highlight from words in the PE version. Therefore, the feature and translation to be judged was highlighted in both MT and PE versions. This meant that we could isolate the formatted words in both versions and obtain the strings to be used by the automatic metrics software.

Nevertheless, we noticed that the mapping of the highlighting to the target text was not always 100% accurate. There was a high number of cases where the highlighted words would not coincide with what a human would consider to be the translation of the -ing word. For instance, as Table 2.20 shows, the word *pending* was correctly translated into the adjective *pendiente* into Spanish, but the RBMT system highlighted two words from the adjacent noun complement as well. Also, we found cases where the RBMT system did not highlight any word in the output.

To run a pending job from the Status view: Para ejecutar una tarea pendiente de la visión de estado
“ Modifying a Backup Selection” on page 1046 “Modificación de una selección de copia de respaldo” en la página 1046

Table 2.20: Examples showing formatting transfer errors by the RBMT system

Such errors would have been corrected by the post-editors. Besides, due to the averaging carried out by the automatic metrics according to the total number of words in the sentences, the scores were bound to change if unnecessary words were included in the strings. In a comparison made between *noisy* data and *clean* data for Spanish, we noted that the correlation results increased by 0.2 for NIST and 0.4 for TER.

We concluded that for a fair comparison of correlation, it was necessary to correct the inaccurate highlighting. For this purpose, based on the evaluation guidelines and the question and answer sessions held with the evaluators, we created guidelines to perform this task manually. The “cleaning” for Spanish and French was performed by the author. Then, the guidelines were refined based on this experience and a native speaker linguist per target language carried out the task for German and Japanese in consultation with the author (see Appendix F for cleaning guidelines).

The guidelines included a short explanation of our motivation for using automatic metrics and how they work. The brief introduction to how automatic metrics work could be seen as biasing the objectivity of the linguists' performance. However, we considered it to be necessary for them to understand the task and the importance of highlighting the correct number of words. Yet, in order to maintain the validity of the task, we did not provide the linguists with the answers from the human evaluation.

The main instruction the guidelines contained was “Highlight all the words in the target language that contain the notions expressed within the -ing constituents such as meaning, tense, aspect, number/gender, agreement, person, etc.” Additionally, the specific scenarios and guidelines previously given to evaluators were included.

The linguists benefited from the guidelines in that the general method was understood. Still, due to the different nature of the languages, new questions emerged and decisions to add or remove the highlight had to be made. For instance, the guidelines informed evaluators that, should the translation of the -ing word include any attachment, this should be highlighted, thus establishing that a word could not be partly highlighted. Whereas this was straightforward for Spanish, French and German, the Japanese case was different. To ensure uniformity for all target languages, we

concluded that we would tokenise Japanese.³¹ Then, this type of decision would always be made based on the boundaries established by the tokenisation program.

Consistency was key to the evaluation and steps were taken to increase the internal consistency for each linguist and across languages. First, the 1,800 sentences were presented to them ordered by -ing category so that the decision to include or exclude a word/character was uniform for each category. Also, they were asked not to pause the task until all the examples pertaining to the category they were working on were completed.

Once the highlighting was fixed for all four target languages, we isolated the formatted words. Note that because we are only focusing on a specific linguistic feature and we isolated particular words, translations of the -ing word spread along the sentence were grouped together. In the example below, the position of *complete* is different in the MT output (incorrect) and the PE version (correct). However, the isolated sequence is the same, which would result in a high automatic score. Therefore, long-distance positioning problems could not be accounted for when using automatic metrics. Yet, local word-ordering and measures of precision and recall would work as per whole sentences.

	Complete sentence	Isolated sequence for evaluation
Source sentence	If your computer is connected to a network, network policy settings might also prevent you from completing this procedure.	completing
MT output	Si su equipo se conecta a una red, la configuración de políticas de red pudo además impedir que usted este procedimiento complete .	que complete
PE version	Si su equipo se conecta a una red, la configuración de políticas de red pudo además impedir que usted complete este procedimiento.	que complete

Table 2.21: Example of automatic metrics' inaccountability of long-distance issues.

³¹ We used ChaSen to tokenise the Japanese sentences <http://chasen.naist.jp/hiki/ChaSen/>

2.4 CHAPTER SUMMARY

This Chapter explained the theoretical and practical elements of the methodologies used to build a representative corpus of IT-domain procedural and descriptive texts from which the -ing words would be extracted for evaluation. A functional scheme consisting of the categories of Adverbials, Characterisers, Progressives, Referentials and Titles (Izquierdo, 2006) was used to classify the -ing words. Next, the evaluation design was considered, taking validity and generalisability into account. We resolved to perform a complementation of human and automatic metrics evaluation. A feature-based evaluation was proposed where the grammaticality and accuracy of -ing words would be assessed by four professional translators. These results would then be compared to the automatic scores obtained by NIST, METEOR, GTM, TER and character-based edit-distance. In the next Chapter, the quantitative and qualitative results from the evaluations will be reported.

CHAPTER 3

CHAPTER 3: DATA ANALYSIS I

This Chapter presents the results obtained from the human and automatic evaluation of -ing words in machine translation into French, German, Japanese and Spanish. Section 1 focuses on the human evaluation. It starts by reviewing the overall results, followed by a more detailed examination of structures generated by the RBMT system. Section 2 reports on the results from the automatic evaluation. It describes the correlations between the human and automatic evaluations in order to test their efficiency in distinguishing between correct and incorrect feature-level machine translation.

3.1 HUMAN EVALUATION

The overall results show that 72-73% of -ing words were correctly handled by SYSTRAN for German, Japanese and Spanish and 52% were correctly handled for French (see Tables 3.1-3.4). It is worth noting that 8% for French, German and Japanese and 5% for Spanish obtained only 2 “correct” votes, which signals good inter-evaluator agreement.

French				
Correct votes	Nº of examples	%		
4	760	42.22	Correct	52.83%
3	191	10.61		
2	148	8.22	Inconclusive	8.22%
1	139	7.72	Incorrect	38.82%
0	562	31.11		

Table 3.1: Overall human evaluation results for French

German				
Correct votes	Nº of examples	%		
4	1,164	64.66	Correct	73.71%
3	163	9.05		
2	147	8.16	Inconclusive	8.16%
1	111	6.16	Incorrect	18.10%
0	215	11.94		

Table 3.2: Overall human evaluation results for German

Japanese				
Correct votes	Nº of examples	%		
4	1,059	58.83	Correct	72.88%
3	253	14.05		
2	154	8.55	Inconclusive	8.55%
1	189	10.50	Incorrect	18.55%
0	145	8.05		

Table 3.3: Overall human evaluation results for Japanese

Spanish				
Correct votes	Nº of examples	%		
4	1,068	59.33	Correct	73.66%
3	258	14.33		
2	99	5.5	Inconclusive	5.50%
1	104	5.77	Incorrect	20.82%
0	271	15.05		

Table 3.4: Overall human evaluation results for Spanish

A closer look at the results per category showed that some categories performed better than other across TL and within each TL (see Table 3.5). Titles and Characterisers were the best performers for German, whereas Referentials obtained the lowest “correct” judgements. For Japanese, Characterisers and Adverbials reached the highest “correct” percentage whereas Progressives obtained a low 40% correct judgements. Spanish performed best for Adverbials and Progressives, whereas Referentials were the weakest. Apart from Progressives, which were the best performers for French with 74% correct translations, the remaining subcategories obtained correctness percentages in the range of the worst performers for the other TLs, with Titles performing worst with 39% correct responses. Similarities across languages also emerged. Characterisers were among the better performers for all languages, whereas Referentials were continuously ranked in the 4th or 5th position. The more in-depth analysis presented in the following sections will shed light on the structures causing problems and the types of errors generated for each subcategory.

	French		German		Japanese		Spanish	
	Category	%	Category	%	Category	%	Category	%
Best performer	progressives	74	titles	80	characterisers	87	adverbials	87
	characterisers	63	characterisers	79	adverbials	75	progressives	82
	adverbials	56	adverbials	68	titles	68	characterisers	75
	referentials	40	progressives	68	referentials	61	titles	64
Worst performer	titles	39	referentials	47	progressives	40	referentials	55

Table 3.5: Classification of best to worst performing category across languages

3.1.1 FRENCH AND SPANISH

In this section a detailed analysis of the performance of each subcategory for French and Spanish will be carried out. Let us start the in-depth analysis by first recalling the -ing word subcategories within each category to then describe the translation structures generated by the RBMT system and compare them against the human evaluation scores. Note that the given percentages might not add up to 100% because only the most salient examples are mentioned and different error types have been combined in particular examples.

3.1.1.1 TITLES

Titles starting with -ing words were first divided into instances where the -ing word appears in initial position and instances where the -ing word is preceded by the preposition *about*. Next, within each group, a distinction was made to reflect the three different locations where the title can be placed: independent titles, titles within quotation marks at the beginning of sentences and titles within quotation marks embedded in a sentence (see Table 3.6).

Subcategory	Nº of examples in sample	Examples
Title_ING_indp	270	Configuring a new deployment
Title_ING_QM1	113	" Editing selection lists" on page 344
Title_ING_QM2	132	For more information, see " Setting default backup options" on page 401.
Title_ABING_indp	22	About applying a ProductName patch
Title_ABING_QM1	12	" About restoring Exchange data from snapshot backups" on page 1250
Title_ABING_QM2	8	For more information, see " About recovering a computer and using the Disaster Recovery Wizard" on page 1505.

Table 3.6: Subcategories within the category of Titles

FRENCH

-ing words directly introducing titles are mainly machine translated into French by SYSTRAN using one of the following structures: nouns, infinitives, gerunds and *de* complementisers. Nouns and infinitives are grammatical solutions for the translation of Titles whereas gerunds are not. *De* complementisers show instances where the gerunds were misanalysed as participial adjectives and therefore translated as modifiers (see Table 3.7 for examples).

French structures for Title_ING		English source	French MT translation ³²
noun		Deleting Backup Selections	Effacer des sélections de sauvegarde
FR ³³	SR	Configuring a new deployment	Configuration d'un nouveau déploiement
43.7%	83.9%		
infinitive		Renaming robotic libraries and drives	Renommer les bibliothèques robotiques et les lecteurs
FR	SR	Deleting a job template for DBA-initiated jobs for Oracle	Effacer un descripteur du travail pour les travaux DBA-lancés pour Oracle
11.48%	48.39%		
gerund		Using the ADBO with the SQL Agent	Utilisant l' ADBO avec le SQL Agent
FR	SR	Installing Backup Exec using the command line (silent mode)	Installant Backup Exec utilisant la ligne de commande (mode silencieux)
20.7%	0%		
<i>de</i> complementiser		Installing and configuring servers for failover or load balancing	Serveurs de installation et de configuration pour la sauvegarde ou l'équilibrage de charge
FR	SR	Backing up database files in a Microsoft cluster	Bases de données de sauvegarde dans une batterie de Microsoft
20.4%	5.5%		

Table 3.7: French translation structures for the subcategory of titles starting with an -ing word

For independent titles, the most recurrent word class in French was the noun with 43.7% representation and 83.9% marked “correct”. Incorrect output seems to arise from terminological inaccuracy. Next, gerunds were generated 20.74% of the time with all examples evaluated as incorrect. *de* complementisers follow closely with 20.37% of the share and all examples, except three real participial adjectives, were evaluated as incorrect. Finally, 11.48% of the examples were turned into infinitives and 48.39% were judged as correct. Note that, in French, titles starting with infinitives are not the norm, despite not being ungrammatical and results reflect this: none of the examples scored 4 points, 48.39% of the examples got 3 points, 38.71% were inconclusive with 2 points and none was classified as incorrect for grammatical reasons (see Table 3.8).³⁴

³² Note that the highlight in the machine translations was automatically placed by the RBMT system and might not represent the exact equivalent or even be present.

³³ FR refers to the frequency rate of the translation structure; SR refers to the success rate of the translation structure.

³⁴ Four examples showing clear terminological inaccuracies were classified as incorrect.

Issues for Title_ING_indp	English source	French MT translation
stylistic issue	Adding a user-defined selection to the User-defined Selections node	Ajouter une sélection définie pour l'utilisateur au noeud Utilisateur-défini de Selections
	Executing queries on agentless target machines using native credentials	Exécuter des requêtes sur les machines cibles agentless utilisant les qualifications indigènes
lexical inaccuracy	Upgrading an existing CASO installation	Évolution d'une installation existante de CASO
	Viewing a device's SCSI information	Visionnement de l'information du SCSI d'un dispositif

Table 3.8: Issues found for the translation of the subcategory of titles starting with an -ing word

Titles within quotation marks at the beginning of sentences were translated into nouns 48.67% of the time with a success rate of 78.18%. *de* complementisers were generated 41.81% of the time and gerunds 34.54% of the time with all examples evaluated as incorrect for both structures. Finally, infinitives were represented 21.82% with a 50% success rate. Terminological inaccuracy and stylistic issues are the source for low scores in this subcategory (see Table 3.9).

Issues for Title_ABING_QM1	English source	French MT translation
stylistic issue	" Adding a duplicate backup template to a policy" on page 446.	« Ajouter un descripteur de sauvegarde double à une politique » à la page 446.
	" Renaming a cascaded drive pool" on page 199	« Renommer un regroupement d'entraînement monté en cascade » à la page 199
terminological inaccuracy	" Copying jobs instead of delegating jobs" on page 830	« Tirages copies au lieu des travaux de délégation » à la page 830
	" Viewing a media server's system properties" on page 203	« Visionnement des propriétés de système d'un serveur multimédia » à la page 203

Table 3.9: Issues found for the translation of the subcategory of titles within quotation marks at the beginning of sentence

Thirdly, titles within quotation marks embedded in sentences were translated mainly as nouns (49.24%) with a success rate of 75.38%; gerunds (23.48%) and *de* complementisers (18.94%) with all examples evaluated as incorrect; and infinitives (5.3%) with a success rate of 14.28%. Similarly to the previous two groups,

terminological accuracy and stylistic issues for infinitives appear to be the source of incorrect output (see Table 3.10).

Issues for Title_ABING_QM2	English source	French MT translation
stylistic issue	For information on adding devices to a device pool, see " Adding devices to a device pool" on page 190.	Pour l'information sur ajouter des dispositifs à un Pool d'appareils, voir le « Ajouter des dispositifs à un Pool d'appareils » à la page 190.
	Refer to " Enabling TSM support" on page 774.	Référez-vous au « Permettre le support de TSM » à la page 774.
terminological inaccuracy	See " Publishing the remote Windows computer to media servers" on page 923.	Voir le « Édition de l'ordinateur Windows distant aux serveurs multimédias » à la page 923.
	If you are upgrading from a previous version of Backup Exec, see " Upgrading from previous versions of Backup Exec" on page 127.	Si vous améliorez d'une version préalable de Backup Exec, voir le « Évolution des versions préalables de Backup Exec » à la page 127.

Table 3.10: Issues found for the translation of the subcategory of titles within quotation marks embedded in a sentence

Not being able to generalise the results for each of the three subcategories of titles starting with the preposition *about* due to the small number of examples in our corpus, we will combine the results and draw conclusions on the titles starting with *about* followed by an -ing word regardless of their location in the text. The preposition *about* is translated as *au sujet de* and the -ing word is mainly (83.33%) translated into an infinitive. This structure is not grammatical, as the French prepositional phrase should be followed by a noun, not an infinitive. However, note that by using the preposition at the beginning of a title, the variation in word class for the -ing word (noun, infinitive and gerund) was reduced (3 examples were translated as nouns, 2 of which were evaluated as incorrect) (see Table 3.11). This means that ways of obtaining consistency by externally manipulating the text exist and will be discussed at a later stage when methods for improvement are considered.

Issues for Title_ABING	English source	French MT translation
infinitives	About restoring Exchange data from snapshot backups	Au sujet de restaurer des données d'Exchange des sauvegardes d'instantané
	" About creating and updating recovery media" on page 1489	« Au sujet de créer et de mettre à jour des medias de reprise » à la page 1489
noun	" About Backing up a DPM server" on page 1473	« Au sujet du Support-vers le haut un serveur de DPM » à la page 1473

Table 3.11: Issues found for the translation of the subcategory of titles starting with *about* followed by an -ing word

SPANISH

Similar to French, the word class used to translate -ing words directly introducing titles are mainly nouns, infinitives, gerunds and *de* complementisers (see Table 3.12). Nouns and infinitives are grammatical options whereas gerunds are not. The MT system generated *de* complementisers to translate -ing words misanalysed as participial adjectives. The frequency with which each of the word classes was generated diverged from the English-French language pair.

Spanish structures for Title_ING	English source	Spanish MT translation
noun FR 14.8%SR 85%	" Redirecting restore jobs for Lotus Domino databases" on page 1311	"Reorientación de las tareas del restablecimiento para las bases de datos de Lotus Domino" en la página 1311
	Adding users to access Information Server	Adición de usuarios al servidor de información del acceso
infinitive FR 49.6%SR 93.3	" Creating selection lists" on page 340	"Crear listas de selección" en la página 340
	" Deleting scheduled jobs" on page 464	"Eliminar tareas programadas" en la página 464
gerund FR 23.7%SR 0%	See " Using Delta File Transfer" on page 1048.	Vea el "Usando transferencia de archivos del delta" en la página 1048.
	Pausing and resuming devices	Interrumpiendo momentáneamente y continuando los dispositivos
<i>de</i> complementiser FR 9.6%SR 11.5%	" Viewing SCSI information for a robotic library" on page 225	"Información del SCSI de la visualización para una biblioteca robótica" en la página 225
	Formatting media in a drive	Soportes del formato en una unidad

Table 3.12: Spanish translation structures for the subcategory of titles starting with an -ing word

The most frequent word class used to translate initial -ing words in independent titles into Spanish was the infinitive (49.63%) with a success rate of 93.28%. Sporadically, SYSTRAN introduced a determiner in front of the infinitive, which is ungrammatical, resulting in incorrect output (see Table 3.13). Gerunds are the second most frequent word class (23.7%) with all examples evaluated as incorrect. Nouns are next with 14.81% of the examples and 85% success. Finally, a few examples of *de* complementisers (9.63%) were generated, with only three real participial adjectives evaluated as correct.

Issues for Title ING indp	English source	Spanish MT translation
determiner + infinitive	Uninstalling from Solaris target machines	El desinstalar de los equipos de destino de Solaris
	Logging on to the management console	El iniciar sesiónse a la consola de administración

Table 3.13: Issues found for the translation of the sucategory of titles starting with an -ing word

Titles within quotation marks at the beginning of sentences were translated mainly into infinitives (55.3%) with a success rate of 94.52%; gerunds were generated 22.73% of the time with all examples evaluated as incorrect; nouns 10.61% of the time with a success rate of 73.33%; and finally, *de* complementisers were incorrectly used to translate 6.82% of the gerunds. Note that similarly to the previous subgroup, the MT system occasionally placed a determiner in front of the infinitive, which is ungrammatical (see Table 3.14). Note that this is a language-specific issue, which did not happen for French.

Issues for Title_ING_QM1	English source	Spanish MT translation
determiner + infinitive	For more information on excluding files from Delta File Transfer, see " Configuring Global Exclude Filters" on page 1064.	Para obtener más información en la exclusión clasifia de transferencia de archivos del delta, ven " El configurar global excluye los filtros" en la página 1064.
	If you want to set commands to run before or after the job, on the Properties pane, under Settings, click Pre/Post Commands and complete the options as described in " Running pre and post commands for restore jobs" on page 508.	Si usted quiere configurar comandos de ejecutarse antes o después de la tarea, en el panel de las propiedades, bajo configuración, haga clic en pre/los comandos del poste y complete las opciones según lo descrito en " El ejecutarse pre y comandos del poste para las tareas del restablecimiento" en la página 508.

Table 3.14: Issues found for the translation of the sucategory of titles within quotation marks at the beginning of sentence

The examples in the third subgroup, titles within quotation marks embedded in sentences, were translated mainly as infinitives (52.21%) with a success rate of 93.22%; nouns (19.47%) with a success rate of 86.36%; and gerunds (18.58%) and *de* complementiser (18.94%) with all examples evaluated as incorrect. Similar to the previous two groups, occasionally a determiner was placed in front of infinitives, resulting in incorrect output (see Table 3.15).

Issues for Title_ING_QM2	English source	Spanish MT translation
determiner + infinitive	" Restoring from SQL filegroup backups" on page 1350	" El restaurar de las copias de respaldo SQL del filegroup" en la página 1350
	" Recovering SQL manually" on page 1371	" El recuperarse SQL manualmente" en la página 1371

Table 3.15: Issues found for the translation of the sucategory of titles within quotation marks embedded in a sentence

The limited data for each subgroup does not allow us to generalise the results for each subgroup within the titles starting with *about* followed by an -ing word. However, we considered it important to report them and we will combine all three subgroups in order to draw conclusions on the titles starting with *about* followed by an -ing word regardless of their location in the text. The preposition *about* is translated as *sobre* and the -ing word is translated into either an infinitive (64.28%) or a noun (16.67%) (see

Table 3.16). Note that the variation of word classes into which the -ing word is translated is significantly reduced by placing a preposition in front of the -ing word. Moreover, the word classes generated in this manner are all correct.³⁵ As happened with the English-French language pair, we observe that it is possible to control the output of the MT system to a certain extent via the source text. Such options will be considered in Chapter 4 when studying ways for improvement.

Issues for Title_ABING	English source	Spanish MT translation
infinitives	About applying a ProductName patch	Sobre aplicar un parche de ProductName
	" About preserving the integrity of the SAP Agent catalog" on page 1425	"Sobre conservar la integridad del catálogo del agente de SAP" en la página 1425
noun	If the RSG resides on a different Exchange server than the databases you are restoring, see " About redirecting Exchange storage group and database restores" on page 1260.	Si el RSG reside en un diferente servidor del intercambio que las bases de datos que usted está restaurando, vea el "Sobre la reorientación de almacenamiento del intercambio agrupe y los restablecimientos de base de datos" en la página 1260.
	About performing a DBA-initiated backup job for Oracle	Sobre la ejecución de una tarea de copia de respaldo DBA-iniciada para Oracle

Table 3.16: Issues found for the translation of the sucategory of titles starting with *about* followed by an -ing word

SUMMARY

Overall for Titles, we observe a better performance for the English-Spanish language pair with a 64% success rate and poorer results for French with only 39% of the examples evaluated as correct. The differences between the languages arise from the fact that *de* complementisers are generated more often in French than in Spanish. Whereas in Spanish both infinitives and nouns are considered correct by evaluators, the infinitives, with a 15.26% of the share, are mostly penalised for French.

-ing words in this category were mainly translated as nouns, infinitives, gerunds and *de* complementisers. One of the difficulties present in titles is the need for the MT system to disambiguate between post-modifying -ing words and gerund-participles.

³⁵ Note that despite all instances being correct, infinitives consistently obtain 3 points whereas nouns obtain 4 points. Evaluators seem to have a preference for nouns over infinitives.

Because titles are not complete sentences and there is a lack of context, this is not always a straightforward task. Yet, note that the RBMT system only generates modifiers (*de* complementisers) for 7.72% of the examples evaluated for Spanish and 19.21% for French.³⁶ It should be mentioned that although -ing words translated as nouns and infinitives are followed by determiners and nouns, those translated as *de* complementisers are always followed by a noun. This leads us to believe that when the -ing word is directly followed by a noun or adjective the disambiguation rules for the RBMT system opt for analysing the -ing word as a modifier rather than a gerund-participle.

Despite the disambiguation between the different -ing word classes being successful, however, all three structures generated by the MT system were not considered correct by evaluators. French prefers nouns whereas Spanish prefers infinitives, although nouns are acceptable. In fact, it should be mentioned that the RBMT system generates these word classes with different frequencies for each particular language pair, which demonstrates attempts at adapting each translation direction to the grammatical requirements and stylistic preferences of the target language.

The RBMT system generates nouns, infinitives and gerunds as translations for -ing words introducing titles. The justification for the MT system generating one or other word class in each example is not evident by looking at the syntax and lexis. -ing words followed by nouns, both common and proper, or determiners are translated into any of the three main options. If we examine the lexical items, we notice that, with one exception for French and two for Spanish, terms are translated into either nouns or infinitives. Moreover, they can all have counterparts translated into Spanish gerunds and alternatively into French present participles. However, not enough examples of each lexical item are available to draw any significant conclusion.

Finally, it is worth mentioning that by analysing titles starting with the preposition *about* followed by an -ing word we discovered that it is possible to obtain a more consistent output. We observed that whereas the -ing words placed in first position in titles could be translated as three different word classes, when preceded by *about*, only infinitives and nouns were generated.

³⁶Not all titles are ambiguous. For instance, initial -ing words followed by determiners are easily identifiable as non-modifiers. However, incompleteness and lack of context still apply.

3.1.1.2 CHARACTERISERS

-ing words functioning as characterisers were classified as pre-modifiers and post-modifiers. The latter, in turn, were divided into reduced relative clauses and nominal and adjectival adjuncts (see Table 3.17). In total, therefore, four subgroups were created for the characterisation function.

Subcategory	Nº of examples in sample	Examples
Char_PR	401	On the product's uninstall warning message, click OK to continue.
Char_POrr	84	The drive containing the Backup-to-disk folder is full.
Char_POnn	49	See your Microsoft Windows documentation for instructions on changing the driver signing policy.
Char_POadj	2	We hope that you have found this document useful in learning more about the product.

Table 3.17: Subcategories within the category of Characterisers

FRENCH

Pre-modifiers were translated into two main structures in French: adjectives (247 examples) and *de* complementisers (125 examples). These two structures covered 92.77% of the pre-modifying -ing words (see Table 3.18), showing high consistency in response from SYSTRAN.

French structures for Char_PR		English source	French MT translation
adjectives		Rules to handle conflicting job start times	Règles pour manipuler des temps de démarrage contradictaires du travail
FR	SR	To list semaphores, use the following command:	Pour mentionner des sémaophores, employez la commande suivante :
61.6%	76.9%		
<i>de</i> complementiser		After defining the cleaning slot, you can set up a cleaning job for the robotic library drive.	Après définition du slot de nettoyage, vous pouvez installer un travail de nettoyage pour le lecteur de bibliothèque robotique.
FR	SR	Select this check box to enable the error-handling rule, or clear the check box to disable the rule.	Choisissez cette case à cocher pour permettre à la règle de erreur- manipulation , ou clair la case à cocher d'invalider la règle.
31.1%	64.8%		

Table 3.18: French translation structures for the subcategory of pre-modifying -ing words

The translation of pre-modifiers as adjectives was evaluated as correct for 76.92% of the examples. Incorrect examples can mainly be attributed to terminological inaccuracy, incorrect positioning of the adjective, that is, the adjective is not modifying the correct head, agreement mistakes between the head and the adjective, and untranslated words (see Table 3.19).

Issues for Char_PR - adjectives	English source	French MT translation
terminological inaccuracy	A third party ISO 9660-compliant CD burning application to burn the IDR-created bootable CD image to a CD.	Une application brûlante CD conforme d'OIN 9660 de tiers pour brûler l'image Différence-crée de CD bootable à un CD.
wrong positioning of adjective	"Copy database" on page 55 lets you copy an existing Backup Exec Database from a computer using a specific name to another computer using the same name.	Le "Copiez la base de données" à la page 55 vous laisse copier un Backup Exec existant Database à partir d'un ordinateur utilisant le même nom.
agreement mistakes between head and modifier	Before reverting a database, Backup Exec deletes all existing database snapshots, including those created with SQL 2005, with the exception of the snapshot used for the revert.	Avant de retourner une base de données, Backup Exec efface tous les instantanés de base de données existante , y compris ceux créés avec SQL 2005, excepté l'instantané utilisé pour le retour.
untranslated words	Use the file versioning feature.	Utilisez le dispositif versioning de fichier.

Table 3.19: Issues found for the translation of the subcategory of pre-modifying -ing words

The translation of pre-modifying -ing words as *de* complementisers was evaluated as correct for 64.8% of the examples. Incorrect examples could be attributed to terminological inaccuracy, particularly *error-handling*, with 25 entries (31.56%), not encoded in the user dictionary as an adjective, and untranslated terms (see Table 3.20). Yet, it is important to note that despite using inaccurate terminology, the grammatical structure was correct. This means that the analysis and generation stages of the system functioned properly, and the error was induced by the lack of a correct term entry in the UD.

Issues for Char_PR – <i>de</i> complementisers	English source	French MT translation
terminological inaccuracy	In some cases this happens because the computer is on the Internet and accessible from within the company's private network, but cannot be located by using just its name or normal browsing methods.	Dans certains cas ceci se produit parce que l'ordinateur est sur les internets et accessible du réseau privé de la compagnie, mais ne peut pas être situé près d'utiliser juste son nom ou méthodes normales de furetage .
	Additional usage details and a grooming function are available in the Usage Details dialog:	Les détails supplémentaires d'utilisation et une fonction de toiletage sont disponibles dans le dialogue d'Usage Details :
	Enter any miscellaneous information for the error-handling rule.	Écrivez n'importe quelle information diverse pour la règle de erreur- manipulation .
untranslated terms	Select Warning entries only to display only entries for warnings.	Choisissez les entrées de Warning pour afficher seulement des entrées pour des avertissements.

Table 3.20: Issues found for the translation of the subcategory of pre-modifying -ing words

Within the post-modifying -ing word subcategory, reduced relative clauses were mainly translated as French present participles (76 examples out of 84), 67.1% of which were evaluated as correct (see Table 3.21). The RBMT system recognised the reduced relative structure in the source and generated a parallel grammatical structure in French, generating a present participle after the head, a valid structure in this target language.

French structures for Char_PRrr		English source	French MT translation
present participles		The drive containing the Backup-to-disk folder is full.	Le lecteur contenant le répertoire de Sauvegarde-à-disque est plein.
FR	SR	Number of allocated media (media belonging to a user media set).	Nombre de medias assignés (medias appartenant aux medias d'un utilisateur réglés).
90.5%	67.1%		

Table 3.21: French translation structures for the subcategory of reduced relative clauses

Examples with low points were mainly instances of reduced relative clauses in the passive form. The MT system replicated the structure generating passive voice reduced relative clauses. Evaluators penalised this structure, as it acquires the meaning of "after doing something", which is inaccurate for this context. Moreover, the progressive aspect introduced by the use of a continuous tense passive structure in English is lost

for French. Although it is not always necessary to make the progressive aspect explicit in French, there are instances where it is necessary to report a precise instruction. The examples with zero points also included instances where the participle modified the incorrect part of the noun phrase (see Table 3.22).

Issues for Char_POrr	English source	French MT translation
passives – progressive aspect lost	Type the password for logging into the system being restored .	Tapez le mot de passé pour enregistrer dans le système étant restauré .
	This class is where the bulk of the data being backed up resides.	Cette classe est où la partie des données étant sauvegardées réside.
	If a Backup Exec for NetWare Servers job is targeted to a drive being used for a Backup Exec for Windows Servers job, the drive appears as reserved.	Si un Backup Exec pour le travail de serveurs NetWare est visé à un lecteur étant utilisé pour un Backup Exec pour le travail de Serveurs Windows, le lecteur apparaît comme réservé.
incorrect head modifier	For more information, see “Restoring the cluster quorum to a Windows 2000 or Windows Server 2003 node running Active Directory to a Microsoft cluster” on page 740.	Pour plus d’information, voir le “Restauration du quorum de batterie Répertoire actif 2003 courant de Windows 2000 ou de Serveur Windows à nœud à une batterie de Microsoft » à la page 740.

Table 3.22 Issues found for the translation of the subcategory of pre-modifying -ing words

Nominal and adjectival adjuncts were the second and third subgroups within the post-modifying -ing words. They modify a noun or an adjective respectively and are placed in front of it joined by a preposition. The -ing words in these structures are required by the joining preposition which, in turn, is dependent on the head the -ing word is modifying (see Table 3.23). Due to the sparse examples of adjectival adjuncts, we will focus on nominal adjuncts only.

French structures for Char_POnn		English source	French MT translation
nominal adjuncts		Selection lists provide a quick and easy way of selecting files that you back up often.	Les listes de sélection fournissent un rapide et une manière simple de choisir les fichiers que vous sauvegardez souvent.
FR	SR	By continuing to run backup jobs on the designated failover node, any further risk of having to restart the jobs again when the failed server rejoins the cluster is avoided.	Par la continuation pour exécuter des jobs de sauvegarde sur le nœud indiqué de sauvegarde, tout autre risque de devoir relancer les travaux de nouveau quand le serveur défaillant rejoint la batterie est évité.
100%	10.2%		

Table 3.23: French translation structures for the subcategory of nominal adjuncts

Nominal adjuncts were translated by a parallel adjunct in French, which is a grammatical structure. However, the preposition and word class following it cannot always be translated using the parallel lexical items of the source language. For instance, *on* might need to be translated as *sur* or *en* depending on the structure. Similarly, *on* + *ing* should not be translated as *sur* + *gerund* but as *sur* + *noun*. We observe that due to the word-by-word translation of this type of structure, only 5 examples out of 49 were evaluated as correct.

The examples evaluated as incorrect contain inappropriate prepositions or, more frequently, inadequate word classes following these prepositions (see Table 3.24).

Issues for Char_POnn	English source	French MT translation
incorrect preposition	For requirements on redirecting this restore	Pour des conditions sur réorienter cette restauration
incorrect word class	Go to “Restoring data by setting job properties” on page 498 for details on selecting restore job properties.	Allez au “Restaurant des données en plaçant des propriétés du travail” à la page 498 pour des détails sur choisir des propriétés du travail de restauration.
	For information on creating jobs with policies, see “Creating jobs using policies” on page 442.	Pour l’ information sur créer des emplois avec des politiques, voir le « Création des emplois utilisant des politiques » à la page 442.

Table 3.24: Issues found for the translation of the subcategory of nominal adjuncts containing -ing words

SPANISH

Spanish results show a behaviour similar to the English-French language pair. Pre-modifiers were translated into adjectives (289 examples) and *de* complementisers (79 examples) and additionally, relative clauses were also generated for 24 examples (see Table 3.25). These three structures covered 97.75% of the pre-modifying -ing words suggesting that high consistency in response is also true for the English-Spanish language pair.

Spanish structures for Char_PR		English source	Spanish MT translation
adjectives		The SAP Agent supports the following SAP DB backup functions:	El agente de SAP admite las funciones de copia de respaldo siguientes del DB de SAP:
FR	SR	To uninstall an existing feature, use the REMOVE command.	Para desinstalar una función existente , utilice el comando del QUITAR.
72.1%	94.12%		
<i>de</i> complementiser		Error messages are displayed in red and warning messages in orange to facilitate troubleshooting.	Los mensajes de error se visualizan en rojo y mensajes de advertencia en naranja para facilitar el localizar averías.
FR	SR	Enter the name of an existing database file from which to copy, or click Browse to navigate to the location of the existing database.	Escriba el nombre de un archivo de base de datos de existencia de el cual copiar, o hacer clic en vaya a naveguen a la ubicación de la base de datos existente.
19.7%	59.5%		
relative clauses		Select this option to configure a schedule for a recurring job, and then click Edit Schedule Details to set the schedule.	Seleccione esta opción para configurar una programación para una tarea que se repite , y después haga clic en editan los detalles de la programación para configurar la programación.
FR	SR	Library sharing prerequisites	Biblioteca que comparte requisitos previos
6%	25%		

Table 3.25: Spanish translation structures for the subcategory of pre-modifying -ing words

The translation of pre-modifying -ing words as adjectives was evaluated as correct in 94.12% of the examples. The examples evaluated as incorrect seem to be due to terminological inaccuracy, incorrect positioning of the adjective, that is, the adjective is not modifying the correct head, agreement mistakes between the head and the adjective, and untranslated words (see Table 3.26).

Issues for Char_PR – adjectives	English source	Spanish MT translation
terminological inaccuracy	A third party ISO 9660-compliant CD burning application to burn the IDR-created bootable CD image to a CD.	Una aplicación ardiente CD obediente de la ISO 9660 del tercero para quemar la imagen CD de arranque IDR-creada a un CD.
wrong positioning of adjective	As soon as the media server is updated, either through a full install, hotfix or Service Pack release, the Desktop Agents will need to be updated in one of the following ways:	Tan pronto como actualicen al servidor de soportes, o a través de un completo instalese, corrección o Service Pack desacopla, los agentes de escritorio necesitará ser actualizado uno de los siguientes de maneras:
agreement mistakes between head and modifier	Note the following items when using continuous protection as part of your backup strategy:	Tenga en cuenta lo siguiente elementos al usar la protección continua como parte de su estrategia de copia de respaldo:
untranslated words	Use the file versioning feature.	Utilice la función versioning del archivo.

Table 3.26: Issues found for the translation of the sucategory of pre-modifying -ing words

The translation of pre-modifying -ing words as *de* complementisers was evaluated as correct in 59.5% of the examples. As with French, a clear cause for this low score was identified. Although the generated structure is valid to modify a head in Spanish, it was observed that 25 examples (31.56%) contained the -ing word *error-handling*. This term was not encoded in the user dictionaries and the MT system proposed an inaccurate translation. Yet, it is important to mention once again that despite using inaccurate terminology, the grammatical structure was correct. This means that the analysis and generation stages were successful, and the error was generated from the lack of a correct term entry in the UD. Some untranslated terms were also found and penalised (see Table 3.27).

Issues for Char_PR – <i>de</i> complementisers	English source	Spanish MT translation
terminology inaccuracy	Enter a new name or change the name for the error-handling rule.	Escriba un nuevo nombre o modifique el nombre para la norma de la error- administración .
untranslated terms	Specify the major number of the agent versioning package that is to be installed on the UNIX target machine.	Especifique el número importante del paquete versioning del agente que debe ser instalado en el equipo de destino de UNIX.

Table 3.27: Issues found for the translation of the sucategory of post-modifying -ing words

Thirdly, the translation of pre-modifying -ing words as relative clauses was evaluated as correct in 25% of the 24 examples. A relative clause is a grammatical structure for modifying a head in Spanish. However, the examples evaluated as incorrect were cumbersome (stylistically inaccurate) or placed the relative clause in a position where it was modifying the wrong head (see Table 3.28).

Issues for Char_PR – relative clauses	English source	Spanish MT translation
stylistic inaccuracy	Profiles can be modified as required to meet the changing needs of user groups.	Los perfiles se pueden modificar como sea necesario para cumplir las necesidades que se modifican de grupos de usuario.
	Backup Exec media server entries in the ralus.cfg file are entered using either the media server name from a naming service provider such as DNS, NIS, or an IP address.	Las entradas del servidor de soportes de Backup Exec en el archivo de ralus.cfg se escriben usando cualquier el nombre del servidor de soportes de un prestatario de servicios que da un nombre a tal como DNS, NIS, o una dirección IP.
incorrect head choice	Support for the DB2 log archiving methods that are known as user exit and VENDOR.	Ayuda para el registro DB2 que archiva los métodos que son conocidos como la salida de usuario y DISTRIBUIDOR.
	The Lotus Domino logging style must be set to archive if you want to back up the transaction logs.	Lotus Domino que registra estilo se debe configurar para archivar si usted quiere hacer copia de respaldo de los registros de transacciones.

Table 3.28 Issues found for the translation of the subcategory of post-modifying -ing words

Within the post-modifying -ing word subcategory, reduced relative clauses were mainly translated as Spanish relative clauses (69 examples). Of these, 78.26% were evaluated as correct (see Table 3.29).

Spanish structures for Char_POrr		English source	Spanish MT translation
relative clauses		The drive containing the Backup-to-disk folder is full.	La unidad que contiene la carpeta del Copia de respaldo-a-disco es completa.
FR	SR	Number of errors occurring since the last cleaning job.	Número de errores que ocurren desde la tarea pasada de la limpieza.
82.1%	78.26%		

Table 3.29: Spanish translation structures for the subcategory of reduced relative clauses

Examples evaluated as incorrect were mainly instances of reduced relative clauses in the passive form. The MT system generated a Spanish relative clause introduced by the relative pronoun *que* in the passive form and in the present tense. Therefore, the progressive aspect introduced by the English structure was lost. Also, some examples appear with terminological inaccuracy (see Table 3.30).

Issues for Char_POrr	English source	Spanish MT translation
passive structures and progressive aspect loss	If you have a Windows NTFS compressed partition, Backup Exec displays the uncompressed byte count of the files being backed up while Windows Explorer displays the compressed byte count of the files on the hard drive.	Si usted tiene una partición comprimida de Windows NTFS, Backup Exec visualiza la cantidad de byte sin comprimir de los archivos que son hecho copia de respaldo des mientras que Windows Explorer visualiza la cantidad de byte comprimida de los archivos en el disco duro.
	dr file contains specific information for the computer being protected , including:	el Dr. archivo contiene la información específica para el equipo que está protegido , incluyendo:
terminology inaccuracy	The product eliminates cumbersome and time-consuming tasks facing database administrators, thereby reducing costs.	El producto elimina las tareas complicadas y laboriosas que hacen frente a los administradores de base de datos, de tal modo reduciendo costos.

Table 3.30: Issues found for the translation of the subcategory of reduced relative clauses

Finally, post-modifying -ing word adjuncts were divided into nominals and adjectivals. Due to the sparse examples of adjectival adjuncts, we will focus on nominal adjuncts only (see Table 3.31).

Spanish structures for Char_POnn		English source	Spanish MT translation
nominal adjuncts		Then, locate and read the Chapter about updating content and components.	Entonces, localice y lea el capítulo sobre actualizar el contenido y componentes.
FR	SR	A method of reducing data to expedite transmission time or storage volume.	Un método de reducir datos para apresurar tiempo de transmisión o el volumen del almacenamiento.
100%	36.7%		

Table 3.31: Spanish translation structures for the subcategory of nominal adjuncts

Nominal adjuncts were translated by a parallel structure in Spanish. However, the preposition and word class following it in the source and in the target languages do not always correspond. Therefore, we see that only 18 examples out of 49 were evaluated as correct (see Table 3.32). Whereas *of* seems to translate correctly as *de* in most of the examples, we observe that *on* is quite problematic. Also, *about* + *ing* seems to be translated as *sobre* + *gerund*, which is ungrammatical in Spanish.

Issues for Char_POnn	English source	Spanish MT translation
incorrect preposition	See your Microsoft Windows documentation for instructions on changing the driver signing policy.	Vea su documentación de Microsoft Windows para las instrucciones en modificar la política de firma del controlador.
incorrect word class	For information on creating device pools, see "Creating device pools" on page 189.	Para obtener información sobre creando a grupos de dispositivo, vea el "Crear a grupos de dispositivo" en la página 189.

Table 3.32: Issues found for the translation of the subcategory of nominal adjuncts

SUMMARY

First of all it is important to note the consistency in the correlation between the source -ing word subcategory and the translation structure generated by the RBMT system. The MT system generated a consistent pattern each time it was faced with a specific -ing word subcategory, although we observed slight variation for pre-modifying -ing words. Yet, this could be seen as an attempt to imitate natural language and its array of possibilities to, in this case, modify a head noun. When diverging structures were generated by the MT system, it was usually the case that the sentences presented problems induced by additional complexity of the context. Consistency means predictability, which is a key asset for improvement through controlled language and post-editing.

The RBMT system translated post-modifying -ing words mainly into two structures for French and three for Spanish: adjectives, *de* complementisers and relative clauses. For Spanish, adjectives were correct 99.3% of the time. *de* complementisers were only correct 59.5% of time. However, note that 23/79 were penalised due to terminological inaccuracy. Relative clauses were correct 25% of the time. For French, the success rate for adjectives was 77% and for *de* complementisers 64.8% excluding the 31.56% classified as incorrect due to the terminological inaccuracy of the term *error-handling*.

By looking at the specific lexical items of each -ing word and the structure it was translated into, the choice of structure does not seem to be random. For instance, *ascending*, *descending*, *existing* and *following* were always translated as adjectives whereas *cleaning*, *monitoring* or *auditing* were translated into different word classes. This suggests that the MT system is tuned to treat the most common participial adjectives as pure adjectives. With regard to the use of *de* complementisers or relative clauses, although no claim can be made due to sparse data, we report that each lexical item was rendered as either one or other structure but not both (with the exception of *filtering*). Therefore, it is possible for a rule to exist which governs the choice of structure for each lexical item.

The main issues regarding these structures and terminological accuracy are that the modifier refers to the incorrect head. For the Spanish relative clauses, stylistic inadequacy leading to potential comprehension problems should be added.

Secondly, reduced relative clauses were translated mainly into relative clauses in both languages. The success rate for this was 78% for Spanish and 67% for French. The incorrect output was due to (1) the relative clause modifying the wrong head, (2) inaccurate passive structures for French and cumbersome passive structures for Spanish (reflex passives are more common), (3) loss of progressive aspect in progressive passive structures and (4) terminological inaccuracy.

Thirdly, nominal adjuncts present a very different pattern. These structures can be translated with a parallel adjunct in French and Spanish. However, the preposition in the target language also depends on the noun/adjective that it is modifying. Similarly, the translation of the -ing word depends on the requirements of the target language preposition. Therefore, when translating these structures, the MT system needs to

know the prepositions that are required for each of the nouns/adjectives and the word class required after each preposition. When this information is available, i.e. coded in the dictionary or in transfer rules, the RBMT system produces a correct output. However, when it is not, a word-for-word translation is produced, which often results in incorrect output. Their success rate in this study is 10.2% for French and 36.73% for Spanish.

3.1.1.3 PROGRESSIVES

-ing words introducing progressive aspect to verbal tenses were categorised depending on the tense (past, present, future), voice (active or passive) and modality (neutral or with modals) of the verbal phrase as shown in Table 3.33.

Subcategory ³⁷	Nº of examples in sample	Examples
Prog_PRACT	108	If you are backing up a SQL server, use this entry.
Prog_PRPAS	26	If this media is password protected and is being cataloged by this system for the first time, enter the password.
Prog_PSACT	1	The Hot-swappable Device Wizard waits until any jobs that were processing are completed.
Prog_PSPAS	1	If failover occurs in the middle of backing up a resource, the media that was being used at the time of the failover is left unappendable and new media will be requested upon restart.
Prog_mod	5	In order to restore SQL databases, SQL must be running ; however, SQL cannot be started unless the master and model databases are present.
Prog_non-inflected	1	ProductName client installation software requires that Microsoft Installer 3.1 be running on client computers before installation.

Table 3.33: Subcategories within the category of Progressives

FRENCH

Active voice present continuous verb structures were translated by the French present simple tense for 93.52% of the examples (see Table 3.34). Of these, 79.2% were classified as correct by evaluators.

³⁷ -ing words introducing progressive aspect in past tenses (active and passive forms), in combination with modals and used with the verb *to be* in the infinitive form will not be discussed in this section due to data sparseness.

French structures for Prog_PRACT		English source	French MT translation
present simple tense		Before installing Backup Exec, be sure that all hardware in the SAN is working and configured properly.	Avant d'installer Backup Exec, soyez sûr que tout le matériel dans le San fonctionne et configuré correctement.
FR	SR	Enter 1 if you are restoring NetWare data and want to restore volume restrictions; otherwise, enter 0.	Écrivez 1 si vous restaurez des données de NetWare et voulez restaurer des restrictions de volume ; autrement, écrivez 0.
93.5%	79.2%		

Table 3.34: French translation structures for the subcategory of active voice present continuous tense

The examples evaluated as incorrect include instances where the progressive aspect required to convey meaning accurately is lost. Terminological inaccuracy also contributes to the low scores (see Table 3.35).

Issues for Prog_PRACT	English source	French MT translation
loss of progressive aspect	Enter 1 to display the percent complete number and gauge while a backup job is processing ; otherwise, enter 0.	Écrivez 1 pour afficher le nombre complet de pour cent et pour le mesurer tandis qu'un travail de sauvegarde traite ; autrement, écrivez 0.
	In a cluster environment, if a job created through BACKINT or RMAN is processing and the node fails over, the job operation does not restart from the point when the node went down.	Dans un environnement de batterie, si un emploi créé par BACKINT ou RMAN traite et l'échouer de noeud plus de, l'exécution du travail ne relance pas du point quand le noeud est descendu.
terminological inaccuracy	(Optional) If you are upgrading by using a new computer, on the computer that runs the embedded database, copy the latest backup .zip file from the \\Program Files\Symantec\Symantec Endpoint Security Manager\data\backup\ directory to the same directory on the new computer.	(Facultatif) si vous améliorez en utilisant un nouvel ordinateur, sur l'ordinateur qui exécute la base de données incluse, copiez le dernier fichier de la sauvegarde .zip du \\ Program Files \ Symantec \ point final Security Manager \ données \ sauvegarde \ répertoire de Symantec au même répertoire sur le nouvel ordinateur.
	By default, Backup Exec detects NetWare servers that are publishing using the TCP/IP protocol.	Par défaut, Backup Exec détecte les serveurs NetWare qui éditent utilisant le protocole TCP/IP.

Table 3.35: Issues found for the translation of the subcategory of active voice present continuous tense

The examples of passive voice present continuous verb structures (26) were all translated using the French passive structure *être + past participle*, where the verb *être* was conjugated in the present simple tense. Whereas 18 of them were evaluated as correct, 6 were considered incorrect. The latter examples include number agreement mistakes and terminological inadequacy (see Table 3.36).

Issues for Prog_PRpas	English source	French MT translation
passive: être + past participle	Drives that are being used are considered online and are included in the number displayed.	Pilote qui sont utilisés sont considérés en ligne et sont inclus dans le nombre affiché.
	Select SCSI Bus Reset if your storage device is being shared on a shared SCSI bus.	Choisissez SCSI Bus Reset si votre périphérique de stockage est partagé sur un bus partagé de SCSI.
terminological inaccuracy	If the cache file is located on a different volume than the volume that is being snapped , Backup Exec uses the following calculations:	Si le fichier de cache est localisé sur un volume différent que le volume qui est cassé , Backup Exec utilise les calculs suivants :
	If multiple source volumes (the volumes to be snapped) are being snapped , then multiple cache files (one for each source volume) are located on the volume you specified (if that volume is not being snapped).	Si des volumes de source multiple (les volumes à casser) sont cassés , alors les fichiers multiples de cache (un pour chaque volume de source) sont localisés sur le volume que vous avez spécifié (si ce volume n'est pas cassé).

Table 3.36: Issues found for the translation of the subcategory of passive voice present continuous tense

SPANISH

The English active voice present continuous tense was translated by SYSTRAN into Spanish using the exact same tense for 88.89% of the examples. From these, 90.62% were evaluated as correct (see Table 3.37).

Spanish structures for Prog_PRACT	English source	Spanish MT translation
present simple tense	If you are using Setaid.ini, the feature is not installed.	Si usted está utilizando Setaid.ini, la función no está instalada.
FR 88.9%	Type the recipient for whom you are configuring the notification.	Escriba al recipiente para quien usted está configurando la notificación.
SR 90.62%		

Table 3.37: Spanish translation structures for the subcategory of active voice present continuous tense

The examples evaluated as incorrect (6) include incorrect uses of reflexive pronouns and an instance of terminological inaccuracy (see Table 3.38). We observed that in a few examples where the reflexive pronoun was introduced this was not required and in a few cases where it was necessary, the MT system did not generate it.

Issues for Prog_PRACT	English source	Spanish MT translation
use/lack of reflexive form	If any errors occur while the macro is executing , no changes or additions are made to the TSM server, and an error notification number appears on the final line of the server console.	Si ocurren algunos errores mientras que la macro está ejecutando , no se hace ningunos cambios o adiciones al servidor de TSM, y un número de la notificación del error aparece en la línea final de la consola del servidor.
	From the Administration Console, ensure you are connected to the Primary Group Server and that you are running in partition management mode.	De la consola de administración, asegúrese que le conecten al servidor primario del grupo y eso que usted se está ejecutando en modo de la administración de la partición.
terminological inaccuracy	An indication that the object size is too large, or that the TSM server is running out of storage space, is if the following errors appear in the Event Log:	Una indicación que el tamaño de objeto es demasiado grande, o que el servidor de TSM se está ejecutando de espacio de almacenamiento, es si los errores siguientes aparecen en el registro de eventos:

Table 3.38: Issues found for the translation of the subcategory of active voice present continuous tense

Examples with passive voice present continuous tenses were translated into Spanish present continuous tense reflexive passives 76.92% of the time, with all examples evaluated as correct. The remaining examples were translated into present continuous tense passives using the auxiliary *ser* or *estar* (see Table 3.39).

Spanish structures for Prog_PRpas		English source	Spanish MT translation
present continuous tense reflexive passives		The media is being loaded and positioned on the target device.	Los soportes se están cargando y se están colocando en el dispositivo de destino.
FR	SR	The name of the media server on which the database is being recovered .	El nombre del servidor de soportes en quien se está recuperando la base de datos.
76.9%	100%		
present continuous tense passives		Select SCSI Bus Reset if your storage device is being shared on a shared SCSI bus.	Seleccione el bus SCSI reajustado si su dispositivo de almacenamiento está siendo compartido en un bus SCSI compartido.
FR	SR	If this media is password protected and is being cataloged by this system for the first time, enter the password.	Si este los soportes son contraseña protegida y están siendo catalogados por este sistema por primera vez, escriba la contraseña.
23.1%	0%		

Table 3.39: Spanish translation structures for the subcategory of passive voice present continuous tense

The examples translated into present continuous tense passives using the auxiliary verb *estar* were evaluated as incorrect as it is an ungrammatical structure (see Table 3.40).

Issues for Prog_PRpas	English source	Spanish MT translation
present continuous tense passives - <i>estar</i>	Consider what type of Windows computer is being protected , the available hardware, and the system BIOS when selecting the type of bootable media to create.	Considere qué tipo de equipo de Windows está estando protegido , el hardware disponible, y el sistema BIOS al seleccionar el tipo de soportes de arranque para crear.

Table 3.40: Issues found for the translation of the subcategory of passive voice present continuous tense

SUMMARY

From the analysis of the translation of active and passive voice present continuous tenses we conclude that the English-French and English-Spanish language pairs diverge in structure. For Spanish, a continuous tense is used whereas for French, on the other hand, a simple tense is generated. Therefore, we see that Spanish explicitly conveys the time and aspect information of the source while French does not display any aspect information.

Vinay and Darbelnet describe how seven forms of the English verbal conjugation correspond to the French present simple (1995: 132). However, they explain, “[t]he progressive forms described in English grammars for foreigners as part of the tense system is in reality an aspect” (ibid: 149). And French has its own means of expressing progressive aspect: the phrase *en train de* or the structure *aller + (en) + participe présent*. Yet, they point out that “usually the context suffices to explain that the action is in progress at the time indicated by the tense” (ibid: 150). We therefore understand that, although possible and grammatical, it is not necessary to make the aspect explicit if it can be deduced from the context.

Overall, the success rate for Spanish was 82% and 74% for French. Yet, room remains for improvement. On the one hand, focus should be placed on the use of the reflexive pronouns when combined with the translation of -ing words for Spanish. On the other hand, French would benefit from clues as to when to make progressive aspect explicit and from the enhancement of subject-participle agreement in passive structures.³⁸

3.1.1.4 ADVERBIALS

-ing words with adverbial function were classified firstly according to the type of adverbial clause they composed: time, manner, purpose, contrast or elaboration. Next, subcategories were created depending on the preposition or subordinate conjunction that preceded the -ing word (see Table 3.41).

Subcategory ³⁹	Nº of examples in sample	Examples
Adv_M_by	111	This is commonly accomplished by using a domain administrator account.
Adv_PU_for	98	Policies provide a method for managing backup jobs and strategies.
Adv_T_when	67	This option is only available when performing full backups.
Adv_T_before	38	If jobs are running, let them finish or cancel them before beginning the installation.

³⁸ We are aware that the evaluators were asked to judge sentences out of context, and therefore, instances penalised due to the lack of progressive aspect might have been correct in context.

³⁹ Subgroups with few examples will not be discussed in this section as no quantitative conclusions can be drawn from their analysis.

Adv_T_after	30	If prompted, reboot the computer after uninstalling Backup Exec.
Adv_M_0	28	The target machine can be configured using the bv-Control for UNIX Configuration wizard.
Adv_M_without	19	If the drive is powered off without ejecting the tape first, this information is lost.
Adv_T_while	14	Select this option to restore files and directories backed up from junction point links while retaining the system's current junction points.
Adv_CD_if	4	Symantec recommends using a range of 25 allocated ports for the remote systems if using Backup Exec with a firewall (see "Using Backup Exec with firewalls" on page 415).
Adv_CT_instead of	3	See "Copying jobs instead of delegating jobs" on page 830.
Adv_T_on	2	On selecting this option you must enter the superuser credentials in the respective su User Name and su Password (required) fields.
Adv_R_0	2	A checkpoint restart option allows backup jobs to continue from the point at which the jobs were interrupted rather than starting the backup over again, making the backups faster and requiring fewer media.
Adv_T_between	1	The amount of time to wait between returning the media from the vault and when it was last written to.
Adv_T_from	1	Select this option to have Backup Exec run the tape in the drive from beginning to end at a fast speed, which helps the tape wind evenly and run more smoothly past the tape drive heads.
Adv_T_0	1	When using Backup Exec's Remote Administrator, specifying drive A:
Adv_T_in the middle of	1	If failover occurs in the middle of backing up a resource, the media that was being used at the time of the failover is left unappendable and new media will be requested upon restart.
Adv_T_in	1	This data can be used in planning for additional device allocation, archiving historical data, and improving performance.
Adv_T_through	1	This wizard guides you through installing the appropriate drivers for the storage hardware connected to your system.
Adv_T_along with	1	Along with backing up the SAP database files, you should do the following:
Adv_E_0	1	The contents of the task pane are dynamic, changing according to the view selected from the navigation bar.

Table 3.41: Subcategories within the category of Adverbials

FRENCH

For French, adverbials of mode with the structure *by + ing* were mainly translated as *en + participe présent*, 91%, with 86.14% success rate. The remaining examples were translated using the structure *par + noun* but none were evaluated as correct (see Table 3.42).

French structures for Adv_M_by		English source	French MT translation
gerundif – <i>en + participe présent</i>		By dynamically allocating devices as jobs are submitted, Backup Exec processes jobs quickly and efficiently.	En assignant dynamiquement des dispositifs comme travaux sont soumis, les travaux de processus de Backup Exec rapidement et efficacement.
FR	SR	Select the following fields by holding down the Ctrl key while you make your selection:	Choisissez les zones suivantes en maintenant la touche ctrl tandis que vous faites votre sélection :
91%	86.1%		
<i>par + noun phrase</i>		You can save time by importing templates that contain all or most of the settings you want to use.	Vous pouvez épargner le temps par l'importation des descripteurs qui contiennent toutes les ou la plupart configurations que vous voulez utiliser.
FR	SR	Complete the configuration of the agentless target machine by navigating to the last panel of the Configuration Wizard.	Terminez la configuration de la machine cible agentless par la navigation au dernier groupe de l'assistant de Configuration.
9%	0%		

Table 3.42: French translation structures for the subcategory of -ing words preceded by *by*

-ing words with an adverbial function preceded by the preposition *for* were mainly translated into the French structures *pour + infinitive* (67.35%) and *pour + noun phrase* (20.41%) with a success rate of 81.82% and 40% respectively. Incorrect output was due to complex context, such as the introduction of negation due to the presence of the particle *No* in the noun phrase following the -ing word, but mainly due to terminological inaccuracies. Note that 9 examples were incorrectly analysed as modifiers and translated as *de* complementisers (see Table 3.43).

French structures for Adv_PU_for		English source	French MT translation
<i>pour</i> + infinitive		Enter one of the following values for recovering the database:	Écrivez une des valeurs suivantes pour récupérer la base de données :
FR	SR	Time limit used for determining Stalled communications status.	Délai utilisé pour déterminer le mode de transmissions de Stalled.
67.3%	81.8%		
<i>pour</i> + det + noun		Use this log for troubleshooting .	Utilisez ce logarithme naturel pour le dépannage .
FR	SR	Requirements for using the Advanced Open File Option	Conditions pour l'usage du fichier ouvert Option d'Advanced
20.4%	40%		

Table 3.43: French translation structures for the subcategory of -ing words preceded by *for*

English grammar allows for the use of implicit subjects by introducing the subordinate clause with a temporal and manner subordinate conjunction followed by an -ing word. In order to be grammatical, however, the subject of the main clause and the subordinate clause must be the same. Instances of this type of structure were found in our corpus and extracted for evaluation. In this case, not only do we focus our analysis on the correct transformation of the -ing word, but also examine whether the information about who the subject is, is conveyed.

The first realisation of this type of structure found in the corpus is *when* followed by an -ing word. It was translated into two structures: the gerund and the subordinate conjunction *quand* followed by noun phrases, infinitives or participles (see Table 3.44). Gerunds were generated 86.57% of the time with only 3.44% evaluated as correct. Note that this structure allows for the subject to be implicit in the target language. The use of *quand* was severely penalised by evaluators with no example evaluated as correct. The issues that arose by the use of this structure were the following: firstly, by using a noun phrase after *quand* the -ing word was turned into the subject of the subordinate clause; secondly, in order to generate a grammatical structure, infinitives cannot follow *quand*; and thirdly, the structure *quand* followed by a participle is a calque of the English structure and is ungrammatical in French.

Regarding the implicit subject question, two conclusions can be drawn. Firstly, in IT user guides it is not always easy to distinguish whether the subject performing an action is the machine or the user (e.g. *Select Display filegroups when creating new backup jobs*). This might be disambiguated by humans when reading the sentence in

context or using logic, but not yet by MT systems. Secondly, there are 26 clear instances where the subjects of the main and subordinate clauses do not refer to the same entity in the source hence, ungrammatical cases, and this was not reflected in the output.

French structures for Adv_T_when		English source	French MT translation
gerundif – <i>en</i> + <i>participe présent</i>		This option is only available when performing full backups.	Cette option est seulement disponible en exécutant de pleines sauvegardes.
FR	SR	When creating a script file, do not include all entries	En créant un fichier script, n'incluez pas toutes les entrées
86.6%	3.4%		
<i>quand</i>		The first decision to make when planning a production installation is to select the database server to use.	La première décision pour faire quand la planification d'une installation de production est de choisir le serveur de base de données pour utiliser.
FR	SR	For example, if c:\junctionpoint is linked to c:\, recursion will occur when attempting to back up c:\junctionpoint, and the backup job will fail.	Par exemple, si c:\junctionpoint est lié à c:\, la récursion se produira quand essayant de sauvegarder c:\junctionpoint, et le travail de sauvegarde échouera.
13.4%	0%		

Table 3.44: French translation structures for the subcategory of -ing words preceded by *when*

Before followed by an -ing word is a second structure within the implicit subject group found in the corpus. Its translation was *avant de* followed by an infinitive with a success rate of 84.21%. Once again, the system generated the same output regardless of whether the subjects of the main and subordinate clauses were the same or not (see Table 3.45).

French structures for Adv_T_before		English source	French MT translation
<i>avant de</i> + infinitive		Display the job summary before creating a job	Affichez le résumé du travail avant de créer un emploi
FR	SR	If jobs are running, let them finish or cancel them before beginning the upgrade.	Si les travaux fonctionnent, laissez-les les terminer ou annuler avant de commencer la mise à niveau.
100%	84.2%		

Table 3.45: French translation structures for the subcategory of -ing words preceded by *before*

A third example of this subcategory is *after* followed by an -ing word. The RBMT system translated this structure into two main structures, making no distinction in the cases where the subject of the main and subordinate clauses were different: *après*

followed by a noun phrase or *après* followed by the verb *avoir* in its infinitival form and the past participle. The first structure was considered correct 88.23% of the time whereas the second was considered correct 72.73% of the time. Note that when using the first structure, the -ing word became the subject of the subordinate clause and when using the second structure, the subject of the subordinate clause was implicit (see Table 3.46).

French structures for Adv_T_after		English source	French MT translation
<i>après</i> + noun phrase		After moving the database, the following occurs:	Après déplacement de la base de données, ce qui suit se produit :
FR	SR	After installing CASO, you can do the following to configure your CASO environment.	Après installation de CASO, vous pouvez faire le suivant pour configurer votre environnement de CASO.
56.7%	88.2%		
<i>après</i> + <i>avoir</i> + past participle		Continue executing the commands that appear after executing the setup.sh command.	Continuez d'exécuter les commandes qui apparaissent après avoir exécuté la commande de setup.sh.
FR	SR	After disabling the Backup Exec 8.x and 9.x Agent for UNIX, proceed with the manual Remote Agent installation.	Après avoir invalidé le Backup Exec 8.x et 9.x Agent pour l'UNIX, procédez à l'installation distante manuelle d'Agent.
36.7%	72.7%		

Table 3.46: French translation structures for the subcategory of -ing words preceded by *after*

The structure *without* followed by an -ing word, which results in an adverbial clause of manner, was translated into French by *sans* followed by an infinitive. This was considered correct for 78.95% of the examples. The incorrect output showed erroneous pronoun choice for reflexive verbs and inaccurate terminology (see Table 3.47).

French structures for Adv_M_without		English source	French MT translation
<i>sans</i> + infinitive		If the drive is powered off without ejecting the tape first, this information is lost.	Si le lecteur est mis hors tension sans éjecter la bande d'abord, cette information est détruite.
FR	SR	You can patch multiple computers at the same time without having to visit each computer individually.	Vous pouvez corriger les ordinateurs multiples en même temps sans devoir visiter chaque ordinateur individuellement.
100%	78.9%		

Table 3.47: French translation structures for the subcategory of -ing words preceded by *without*

With only 14 examples evaluated, *while* followed by an -ing word is the last structure of the implicit subject type we discuss here. Note that *while* in English can indicate that the two clauses it is joining happen simultaneously, or that they show contrast. Within the 14 examples, 2 were of the latter case. We observe that the MT system recognised the difference and generated different French structures to convey the different meanings. On the one hand, simultaneity was conveyed by *tout* followed by a gerund which was considered correct for half the examples. On the other hand, contrast was translated as *alors que* followed by an infinitive, which was evaluated as incorrect because, even if the conjunction was correct, the word class following it should have been a noun (see Table 3.48).

French structures for Adv_T_while		English source	French MT translation
<i>tout</i> + gerund		Select this option to restore files and directories backed up from junction point links while retaining the system's current junction points.	Choisissez cette option pour restaurer des fichiers et des répertoires sauvegardés des liens de point de jonction tout en maintenant les points de jonction actuels du système.
FR	SR	While using DLO, you can suppress dialogs by checking the Don't show me this message again check box.	Tout en utilisant DLO, vous pouvez supprimer des dialogues par le contrôle ne m'affichez pas cette case à cocher de message de nouveau.
85.7%	50%		
<i>alors que</i> + infinitive		Adding a media server to the list expands the number of media servers that can back up the Macintosh computer, while deleting a media server removes it from the list of media servers to which it advertises itself.	Ajouter un serveur multimédia à la liste augmente le nombre de serveurs multimédias qui peuvent sauvegarder le Mac, alors qu'effacer un serveur multimédia le retire de la liste de serveurs multimédias auxquels elle s'annonce.
FR	SR		
14.3%	0%		

Table 3.48: French translation structures for the subcategory of -ing words preceded by *while*

Finally, we examine the results for -ing words with adverbial function not preceded by prepositions or subordinate conjunctions. We observed that all examples were translated into French present participles and they were all evaluated as incorrect (see Table 3.49). Note that from the 28 examples, 25 were composed of the -ing word *using*. Options for correct translations would have been the use of French gerunds (*en* + *participe présent*), the phrase *grâce à*, or the prepositions *avec* or *par*.

French structures for Adv_M_0		English source	French MT translation
participe présent		Select this option to install remotely using all of the installation options that are installed on the local computer.	Choisissez cette option pour installer à distance utilisant toutes les options d'installation qui sont installées sur l'ordinateur local.
FR	SR	The target machine can be configured using the bv-Control for UNIX Configuration wizard.	La machine cible peut être configurée utilisant la BV-Control pour l'assistant d'UNIX Configuration.
100%	0%		

Table 3.49: French translation structures for the subcategory of -ing words preceded by \emptyset

SPANISH

For Spanish, the structure *by* followed by an -ing word was translated into gerunds 98.20% of the time. From this, 97.25% were evaluated as correct, showing little room for improvement (see Table 3.50).

Spanish structures for Adv_M_by		English source	Spanish MT translation
gerund		Users can be identified by using the following options:	Los usuarios pueden ser identificados usando las siguientes opciones:
FR	SR	Otherwise, Available Capacity is calculated by subtracting Bytes Written from Total Capacity.	Si no, la capacidad disponible es calculada restando los bytes escritos de capacidad total.
98.2%	97.2%		

Table 3.50: Spanish translation structures for the subcategory of -ing words preceded by *by*

The structure *for* followed by an -ing word was mainly translated as *para* followed by an infinitive (87.75%). 90.70% of the examples were evaluated as correct and none as incorrect (see Table 3.51).

Spanish structures for Adv_PU_for		English source	Spanish MT translation
<i>para</i> + infinitive		Select a backup method for backing up eDirectory data.	Seleccione un método de copia de respaldo para hacer copia de respaldo de datos eDirectory.
FR	SR	Enter one of the following values for recovering the database:	Escriba uno de los siguientes los valores para recuperar la base de datos:
87.7%	90.7%		

Table 3.51: Spanish translation structures for the subcategory of -ing words preceded by *for*

Maintaining the same focus as for the French analysis, we now discuss the structures that allow implicit subjects. The most frequent is *when* followed by an -ing word. This was translated into Spanish using *al* followed by an infinitive 91.04% of the time. Evaluators judged 95.08% of them as correct. Similarly to what happened in French, for 6 examples the MT system generated the subordinate conjunction *cuando* followed by either a noun phrase or an infinitive. No clear lexical or syntactic clues can be reported for this behaviour (see Table 3.52). These examples were incorrect. Note that whereas the former structure is impersonal, the latter required an appropriate identification of the subject of the subordinate clause.

Spanish structures for Adv_T_when		English source	Spanish MT translation
<i>al</i> + infinitive		Also, when selecting objects from the mailbox tree, all objects are displayed as messages.	Además, al seleccionar objetos del árbol del buzón, todos los objetos se visualizan como mensajes.
FR	SR	Only full backups are supported with the Remote Agent when using the Backup Wizard.	Solamente las copias de respaldo completas se admiten con Remote Agent al usar al Asistente de copia de respaldo.
91%	95.1%		
<i>cuando</i> + noun phrase/infinitive		The first decision to make when planning a production installation is to select the database server to use.	La primera decisión para hacer cuando la planificación de una instalación de producción es seleccionar al servidor de base de datos para utilizar.
FR	SR	The default behavior when deleting a message from a mail archive may differ depending on the mail application.	El comportamiento predeterminado cuando eliminar un mensaje de un archivo de almacenamiento del correo puede diferenciarse dependiendo de la aplicación del correo.
9%	0%		

Table 3.52: Spanish translation structures for the subcategory of -ing words preceded by *when*

For *before* followed by an -ing word, the RBMT system generated the impersonal structure *antes de* followed by an infinitive. This was correct for 94.74% of the examples, with the incorrect output being due to terminological inaccuracy and incorrect use of pronouns (see Table 3.53).

Spanish structures for Adv_T_before		English source	Spanish MT translation
<i>antes de</i> + infinitive		Display the job summary before creating a job	Visualice el resumen de la tarea antes de crear una tarea
FR	SR	Before installing DLO, review "Before you install" on page 986.	Antes de instalar el DLO, revise el "Antes de instalar" en la página 986.
100%	94.7%		

Table 3.53: Spanish translation structures for the subcategory of -ing words preceded by *before*

The structure *after* followed by an -ing word was translated into *después de* followed by an infinitive 93.33% of the time. All examples were evaluated as correct. One instance was translated into *después de* followed by a substantive subordinate clause introduced by *que*, which was also correct (see Table 3.54). As for the source, the translation proposed by the RBMT system is impersonal and therefore the subject of the subordinate clause is not made explicit.

Spanish structures for Adv_T_after		English source	Spanish MT translation
<i>después de</i> + infinitive		After checking the check box, type the destination database name.	Después de activar la casilla de selección, escriba el nombre de base de datos del destino.
FR	SR	Continue executing the commands that appear after executing the setup.sh command.	Continúe ejecutando los comandos que aparecen después de ejecutar el comando de setup.sh.
93.33%	100%		

Table 3.54: Spanish translation structures for the subcategory of -ing words preceded by *after*

The structure *without* followed by an -ing word was translated into *sin* followed by an infinitive (73.78%) or *sin* followed by a noun phrase (21.05%). The former structure obtained an 85.71% success rate, whereas for the latter no examples were evaluated as correct (see Table 3.55). Note that both options are impersonal and therefore the subordinate subject is made implicit.

Spanish structures for Adv_M_without		English source	Spanish MT translation
<i>sin</i> + infinitive		You can patch multiple computers at the same time without having to visit each computer individually.	Usted puede aplicar un parche a los equipos varios al mismo tiempo sin tener que visitar cada equipo individualmente.
FR	SR	After a snapshot is created, the primary data can continue being modified without affecting the backup operation.	Después de que se cree una instantánea, los datos primarios pueden continuar siendo modificada sin afectar a la operación de copia de respaldo.
73.8%	85.7%		

<i>sin</i> + noun phrase		If the drive is powered off without ejecting the tape first, this information is lost.	Si la unidad se acciona apagado sin la expulsión de la cinta primero, se pierde esta información.
FR	SR	You can integrate Lotus Domino database backups with regular server backups without separately administering them or using dedicated hardware.	Usted puede integrar las copias de respaldo de base de datos de Lotus Domino con las copias de respaldo regulares del servidor sin por separado la administración de ellas o usar el hardware dedicado.
21%	0%		

Table 3.55: Spanish translation structures for the subcategory of -ing words preceded by *without*

As we mentioned when analysing the French results, *while* in English usually conveys simultaneity but it is also used to convey contrast. Each of these meaning has a different rendering in Spanish. Whereas simultaneity is conveyed with *mientras*, contrast is conveyed with *mientras que*. We observed that this distinction was not made by the MT system and all examples were generated using the conjunction for contrast. Yet, this structure obtained a success rate of 50% and only 1 example was evaluated as incorrect (see Table 3.56).

Spanish structures for Adv_T_while		English source	Spanish MT translation
<i>mientras que</i> + subordinate clause verb		Displays information detailing what has occurred while running the Command Line Applet and the specified option.	Visualiza el detalle de la información qué ha ocurrido mientras que ejecuta Command Line Applet y la opción especificada.
FR	SR	While using DLO, you can suppress dialogs by checking the Don't show me this message again check box.	Mientras que usa el DLO , usted puede suprimir cuadros de diálogo activando no me muestra esta casilla de selección del mensaje de nuevo.
100%	50%		

Table 3.56: Spanish translation structures for the subcategory of -ing words preceded by *while*

Finally, we examine the adverbial clauses of manner introduced by an -ing word. All the -ing words were translated into gerunds. With the exception of one example, they were all evaluated as correct (see Table 3.57).

Spanish structures for Adv_M_0		English source	Spanish MT translation
gerund		Oracle servers and SAP databases cannot be restored using IDR.	Los servidores de Oracle y las bases de datos de SAP no pueden ser restaurados usando el IDR.
FR	SR	Backup Exec functions and utilities that you can run using the Command Line Applet include:	Las funciones y las utilidades de Backup Exec que usted puede ejecutar usando Command Line Applet incluyen:
100%	96.4%		

Table 3.57: Spanish translation structures for the subcategory of -ing words preceded by ø

SUMMARY

Overall, the success rate for the category is 87% for Spanish and 56% for French. This difference can be explained to some extent by observing the consistency with which translation structures were generated for each language. The consistency in the structures generated for the -ing word subcategories is high as the word class into which the -ing word is translated depends on the preposition/subordinate conjunction preceding it. Spanish benefits from this particularly, with a high percentage of examples for each subcategory only being translated into one target structure.

Regarding the structures allowing for implicit subjects, it should be noted that source instances where the subject of the main and subordinate clauses did not refer to the same entity were found. Evaluators were not concerned with this as SYSTRAN mainly generated impersonal structures for both target languages, although the same grammatical principle stands for the target languages. This means that the MT system did not have to identify the subject of the subordinate clause in most of the cases.

REFERENTIALS

-ing words with referential function were classified as gerundial nouns, objects of catenative verbs, objects of prepositional verbs, comparatives and objects of phrasal verbs (see Table 3.58).

Subcategory ⁴⁰	Nº of examples in sample	Examples
Ref_nn	66	The files gathered contain detailed information regarding installation, diagnostics, and error reporting .
Ref_cat	38	The product supports querying target SQL Servers in an untrusted domain.
Ref_prepV	24	You can place a scheduled job on hold to prevent the job from running .
Ref_comp	11	Symantec recommends performing redirected restore of corrupt files rather than restoring to the original location.
Ref_phrV	2	This could happen, for example, when the desktop user logs in using a local or cross-domain account.

Table 3.58: Subcategories within the category of Referentials

FRENCH

-ing words functioning as nouns were translated into French using nouns 84.85% of the time. This was successful 55.36% of the time (see Table 3.59).

French structure for Ref_nn		English source	French MT translation
noun		The default setting is all local drives are published.	Le paramètre par défaut est tous les lecteurs locaux sont édités.
FR	SR	Error messages are displayed in red and warning messages in orange to facilitate troubleshooting .	Des messages d'erreur sont affichés dans les messages rouges et d'avertissement dans l'orange pour faciliter le dépannage .
84.8%	55.36%		

Table 3.59: French translation structures for the subcategory of -ing words functioning as nouns

Examples evaluated as incorrect presented terminological inaccuracy. Another frequent issue appeared to be the difficulty in identifying the correct head-modifier relationship within complex noun phrases. Also, the noun preceding the -ing word was analysed as an infinitive in a few examples (see Table 3.60).

⁴⁰ -ing words in comparative structures and as objects of phrasal verbs will not be discussed in this section as no quantitative conclusions can be drawn from their analysis.

Issues for Ref_nn	English source	French MT translation
terminological inaccuracy	Click Configure Logging .	Le clic configurent l' enregistrement .
	See "File Grooming " on page 1165 for additional information.	Voir le « Classez le toiletage » à la page 1165 pour les informations complémentaires.
complex noun phrase	If the If closed with X seconds setting is selected in Backup Exec, (under Advanced node in the Backup Job Properties pane) its value should not exceed the Communications Timeout setting on the IBM TSM server.	Si si fermé avec des secondes de X l' établissement est choisi dans Backup Exec, (sous le noeud d'Advanced dans le volet de sauvegarde de Job Properties) que sa valeur si dépasse la configuration de Communications Timeout sur le serveur d'IBM TSM.
	A policy for this example staging would include a backup template to back up the data to disk for the 28 days, a duplicate backup set template to copy the data from the original disk to the second disk, and another duplicate backup set template to copy the data from the second disk to the tape.	Une politique pour cet échafaudage d'exemple inclurait un descripteur de sauvegarde pour sauvegarder les données au disque pour que les 28 jours, un descripteur réglé de sauvegarde double pour copier les données à partir du disque initial au deuxième disque, et un descripteur réglé différent de sauvegarde double copie les données à partir du deuxième disque à la bande.
noun / infinitive	See "File Grooming " on page 1165 for additional information.	Voir le « Classez le toiletage » à la page 1165 pour les informations complémentaires.
	Supports spanning of backup sets from one piece of media to another.	Supporte l' enjambement des positionnements de sauvegarde de l'une seule pièce des medias à l'autre.

Table 3.60: Issues found for the translation of the subcategory of -ing words functioning as nouns

-ing words functioning as objects of catenative verbs were mainly translated as infinitives (65.79%). 23.68% were translated as present participles. Whereas the use of infinitives was correct for some structures such as *éviter de*, *recommander de* or *risquer de* to mention a few, others such as *inclure*, *empêcher* or *démarrer*, for instance, need to be followed by a noun phrase (see Table 3.61). The use of present participles, which is the parallel structure of the source, was evaluated as incorrect for all examples.

French structures for Ref_cat	English source	French MT translation
éviter de + infinitive	For example, you may want to avoid selecting entire drives for backup or synchronization.	Par exemple, vous pouvez vouloir éviter de choisir les lecteurs entiers pour la sauvegarde ou la synchronisation.
recommander de + infinitive	Symantec recommends using fully qualified computer names.	Symantec recommande d'utiliser entièrement - des noms d'ordinateur qualifiés.
inclure + infinitive	Major changes include adding , removing, or otherwise modifying hard drives or partitions, file systems, configurations, and so forth.	Les principaux changements incluent ajouter , retirant, ou les disques durs ou les partitions autrement de modification, systèmes de fichiers, configurations, et ainsi de suite.
considérer + participe présent	Consider using differential backups when only a relatively small amount of data changes between full filegroup backups, or if the same data changes often.	Considérez utilisant les sauvegardes différentielles quand seulement un peu de données relativement change entre de pleines sauvegardes de filegroup, ou si les mêmes données changent souvent.

Table 3.61: French translation structures for the subcategory of -ing words functioning as objects of catenative verbs

From the 24 examples of -ing words functioning as objects of prepositional verbs, we observed that 20 were translated into infinitives. This was the case for *baser sur*, *protéger contre*, *s'inquiéter de*, or most examples of *empêcher de*. The verb *aider*, a translation of *aid in* and *assist in*, and the verb *continuer* were translated using nouns or gerunds. Note that only 7 examples were evaluated as correct. Only 5 out of the 13 examples with *empêcher de*, the example with *aider à* and the example with *continuer* were evaluated as correct. Examples evaluated as incorrect presented terminological inaccuracy as well as the need for the -ing word to be translated as a noun phrase instead of an infinitive (see Table 3.62).

French structures for Ref_prepV	English source	French MT translation
protéger contre + infinitive	Caution This option is not recommended because it does not protect data from being overwritten.	L'option de This d'attention n'est pas recommandée parce qu'elle ne protège pas des données contre être recouvert.
continuer + noun	In the second by Control for Microsoft SQL Server message, click OK to delete all the deployed files and continue with uninstalling .	En deuxième BV Control pour le message de Serveur SQL de Microsoft, cliquez sur l'OK pour effacer tous les fichiers déployés et pour continuer la désinstallation .
empêcher de + infinitive	Select Disable network backup to prevent users from backing up to the network user data folder.	Choisissez la sauvegarde de réseau de Disable pour empêcher des utilisateurs de sauvegarder au répertoire de données d'utilisateur du réseau.
empêcher de + infinitive	If your computer is connected to a network, network policy settings might also prevent you from completing this procedure.	Si votre ordinateur est connecté à un réseau, les configurations de politique de réseau pourraient également vous empêcher de remplir ce procédé.

Table 3.62: French translation structures for the subcategory of -ing words functioning as objects of prepositional verbs

SPANISH

For Spanish, -ing words functioning as nouns were translated using nouns 75.76% of the time, with a success rate of 78% (see Table 3.63). The remaining examples were translated into infinitives (6.06%), infinitives preceded by determiners (6.06%) and relative clauses (6.06%).

Spanish structures for Ref_nn	English source	Spanish MT translation
nouns	Questions regarding product licensing or serialization	Preguntas con respecto la concesión de licencia o a la serialización del producto
FR 75.8%	SMTP email or phone text messaging	Envío de mensajes de texto del correo electrónico o del teléfono del SMTP
SR 78%		

Table 3.63: Spanish translation structures for the subcategory of -ing words functioning as nouns

Examples evaluated as incorrect presented mainly terminological inaccuracy. Infinitives were successful 50% of the time, whereas infinitives preceded by determiners and relative clauses were incorrect (see Table 3.64).

Issues for Ref_nn	English source	Spanish MT translation
noun	Set time for automatic clearing of alert or disable automatic alert clearing	La hora determinada para el claro automático de la alerta o deshabilita el claro alerta automático
	Click Ascending to group the information in ascending order or click Descending to group the information in descending order.	Haga clic en la ascensión para agrupar la información en orden ascendente o para hacer clic en el descenso para agrupar la información en orden descendente.
infinitive determiner + infinitive	This product is protected by copyright and distributed under licenses restricting copying , distribution, and decompilation.	Este producto está protegido por el copyright y distribuido bajo las licencias que restringen copiar , la distribución, y la descompilación.
determiner + infinitive relative clause	Error messages are displayed in red and warning messages in orange to facilitate troubleshooting .	Los mensajes de error se visualizan en rojo y mensajes de advertencia en naranja para facilitar el localizar averías .
relative clause infinitive	Automated restore selections and options checking , which tests the validity of your current SQL Server restore selections and job options before the restore job runs.	Automatizado restaure las selecciones y las opciones que activan , que prueba la validez de su SQL Server actual restauran selecciones y opciones de la tarea antes de que la tarea del restaurar se ejecute.

Table 3.64: Issues found for the translation of the subcategory of -ing words functioning as nouns

The word class into which -ing words functioning as objects of catenative verbs were translated was diverse. For instance, the objects of *evitar*, *considerar*, *impedir* and *comenzar* were translated as infinitives and evaluated as correct. Others, such as *detener*, *recomendar*, for instance, were translated as an infinitive preceded by a determiner, which was evaluated as incorrect. The -ing object of *continuar* was translated by gerunds and was correct according to the evaluators. It is important to note that incorrect output was also generated due to terminological inaccuracy and issues related to contexts with additional linguistic features known to pose problems for MT systems (see Table 3.65). Overall, 42.11% of the examples were evaluated as correct.

Issues for Ref_cat	English source	Spanish MT translation
infinitive	Consider running full backups of mailboxes or public folders on a regular basis.	Considere ejecutar las copias de respaldo completas de buzones o de carpetas públicas sobre una base regular.
determiner + infinitive	Symantec recommends storing the catalogs on the central administration server.	Symantec recomienda el almacenar de los catálogos en el servidor de la administración central.
gerund	After checking the return codes, Backup Exec continues processing according to selections you made for running the pre- and post-commands.	Después de activar los códigos del devolver, Backup Exec continúa procesando según selecciones que usted hizo para ejecutar el pre- y los poste-comandos.
terminological inaccuracy	After you have completed adding (or removing) the SQL Servers, click Next.	Una vez que han completado agregando (o quitando) a los servidores SQL, hacen clic en después.
complex context	Backup Exec starts collecting the true image restore information beginning with the next full or incremental backup run by the policy after the option is enabled.	El comienzo de Backup Exec que recopila la imagen verdadera restaura la información comenzando con la copia de respaldo completa o incremental siguiente ejecutada por la política después de que se habilite la opción.

Table 3.65: Issues found in the translation of the subcategory of -ing words functioning as objects of catenative verbs

24 examples were evaluated for -ing words functioning as objects of prepositional verbs. We observed that 13 examples were composed of the prepositional verb *prevent from*. The translation was correctly generated as *impedir que* followed by the verb in the subjunctive form. For the rest of the prepositional verbs, the -ing word was translated into an infinitive in Spanish. Of these, only two examples were evaluated as correct (see Table 3.66). Note that the errors were due to the terminological choice of the prepositional verb, preposition or word class for the translation of the -ing word.

Issues for Ref_prepV	English source	Spanish MT translation
prevent from	The only way to prevent users in a profile from backing up a specific folder is to uncheck this option.	La única manera de impedir que los usuarios en un perfil hagan copia de respaldo de una carpeta específica es desactivar esta opción.
terminological inaccuracy	The ActiveAdmin function is about using the ActiveAdmin privilege to edit a user record of the Users data source.	La función de ActiveAdmin está sobre usar el privilegio de ActiveAdmin de editar un expediente de usuario del origen de datos de los usuarios.
preposition inaccuracy	The media server contains media and device databases designed to simplify the process of organizing and allocating storage devices attached to your media server and to aid in preventing media from being accidentally overwritten.	El servidor de soportes contiene los soportes y las bases de datos de dispositivos diseñados para simplificar el proceso de ordenar y de asignar los dispositivos de almacenaje asociados a su servidor de soportes y a la ayuda en impedir que los soportes sean sobrescritos accidentalmente.
word class error	A suite of pre-defined queries assist you in identifying key issues such as database integrity, security, and permissions tracking.	Un conjunto de consultas predefinidas le asiste en identificar las cuestiones claves tales como integridad de base de datos, seguridad, y seguimiento de los permisos.

Table 3.66: Issues found in the translation of the subcategory of -ing words functioning as objects of prepositional verbs

SUMMARY

This category, covering -ing words with referential function, is the class with the poorest results, with a success rate of 55% for Spanish and 40% for French. From the analysis we can outline the main reasons for this low score. Firstly, we observed that gerundial nouns were treated by the RBMT system as pure nouns and translated into nouns. The main errors for these examples were the same as for pure nouns, that is, terminological inaccuracies and head-modifier dependency mistakes. Therefore, we argue that better performance might be obtained if the terms were encoded in the UD.

Secondly, when examining -ing words functioning as objects of catenative or prepositional verbs, we noticed that the word class into which the -ing word should be translated depended on the main verb. In the cases where the MT system identified the structures and recognised them as catenatives or prepositional verbs, the generation

was correct. However, when not enough information on the required preposition or word class was found in the engine, the system generated infinitives or structures parallel to the source, usually resulting in incorrect output.

3.1.2 JAPANESE AND GERMAN

Let us now focus on the analysis of Japanese and German. As mentioned in Chapter 2, the analysis for these TLs was performed by native speaker linguists. In order to reduce effort and time, the analysis focused on incorrect examples only. This section, therefore, reviews the -ing subcategories for which incorrect examples were found. In particular, it describes the subcategories for which more than 10 examples were judged to be incorrect to report the tendency for the errors generated by the MT system.

3.1.2.1 JAPANESE

From the 1,800 -ing words evaluated, 319 were classified as incorrect for Japanese (17.72%). A closer examination showed that errors were spread across 20 subcategories. We describe the performance for each category by observing the errors generated by the RBMT system.

TITLES

The category of Titles included 557 examples, out of which 117 were classified as incorrect for Japanese (21%). In correlation with their frequency rates, independent titles were the subcategory which performed worst, followed by titles within quotations marks embedded within a sentence or at the beginning of sentence.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Title_ING_indp	65	Creating media vaults	メディアを作成することは ボルト処理します
Title_ING_QM2	28	See " Setting defaults for managed media servers" on page 809.	「設定は管理対象メディア サーバーのためにデフォ ルトします」を p.809 の参 照して下さい。
Title_ING_QM1	23	" Running Backup Exec Utility services tasks" on page 24	「連続した Backup Exec の ユーティリティサービスは 任せます」 p.24 の
Title_ABING_indp	1	About troubleshooting DB2	トラブルシューティング DB2 について

Table 3.67: Subcategories of the category Titles with incorrect examples

65 examples were classified as incorrect for titles beginning with an -ing word head. Three main errors emerged: the gerund-participle was mistranslated as a participial adjective for 21 instances, on 20 occasions it was incorrectly translated as an adverbial of mode introduced by the preposition *by*, and in 11 instances the -ing word was translated as the subject of the following clause, which includes a plural noun or a past participle functioning as an adjective analysed as a verb. Additional errors include dependency parsing errors, particularly when the titles include coordination and terminological inaccuracies.

Error types for Title_ING_indp	Examples	
	English source	Japanese MT translation
adverbial of mode introduced by <i>by</i> ----- FR ⁴¹ – 30.8%	Using the Administration Console	Administration Console を使用して
subject ----- FR – 17%	Searching for files to restore	有無を検索することは復元するためにファイルします
modifier ----- FR – 32.3%	Setting default options for reports	レポートのための設定のデフォルトオプション

Table 3.68: Error types found in the translation of the subcategory of -ing words at the beginning of title

28 examples were classified as incorrect for titles within quotations marks embedded in the sentences. Three main errors emerged: in 9 cases it was incorrectly translated as an adverbial of mode introduced by the preposition *by*, the gerund-participle was mistranslated as a participial adjective for 7 instances, and 4 instances display the translation of the -ing word as the subject of the following clause, which includes a plural noun or a past participle functioning as an adjective analysed as a verb. Additional errors include dependency parsing errors, the introduction of progressive aspect, and terminological inaccuracies.

⁴¹ FR refers to the frequency rate of the translation errors.

Error types for Title_ING_QM2	Examples	
	English source	Japanese MT translation
adverbial of mode introduced by <i>by</i> ----- FR – 32.1%	See " Using Delta File Transfer" on page 1048.	「差分ファイルの転送を使用して」を p.1048 の参照して下さい。
subject ----- FR – 14.3%	See " Setting defaults for managed media servers" on page 809.	「設定は管理対象メディアサーバーのためにデフォルトします」を p.809 の参照して下さい。
modifier ----- FR – 25%	To move a user to a new location, see " Moving Desktop Agent Users to a new Network User Data Folder" on page 1078.	新しい場所にユーザーを移動するためには、「新しい Network のユーザーデータのフォルダへの Users 移動 Desktop Agent」を p.1078 の参照して下さい。

Table 3.69: Error types found in the translation of the subcategory of -ing words embedded within quotation mark embedded in sentences

Similarly to the previous subcategories, titles within quotation marks at the beginning of a sentence also displayed the same three main errors. The gerund-participle was mistranslated as a participial adjective in 10 instances, on 6 occasions it was incorrectly translated as an adverbial of mode introduced by the preposition *by*, and in 3 instances the -ing word was translated as the subject of the following clause, which includes a plural noun or a past participle functioning as an adjective analysed as a verb. Stylistically poor renderings make up for the remaining 4 instances.

Error types for Title_ING_QM1	Examples	
	English source	Japanese MT translation
adverbial of mode introduced by <i>by</i> ----- FR – 26%	" Using Alternate Credentials for the Desktop Agent"	「Desktop Agent のための代替のクレデンシャルを使用して」
subject ----- FR – 13%	" Excluding dates from a schedule" on page 388	「除くことはスケジュールさかのぼります」 p.388 の
modifier ----- FR – 43.5%	" Setting backup options for SQL" on page 1327.	「SQL のための設定のバックアップオプション」 p.1327 の。

Table 3.70: Error types found in the translation of the subcategory of titles starting with -ing words embedded within quotation marks

PROGRESSIVES

Two subcategories from the Progressives category were judged as incorrect: present active and passive voice subcategories, with 52 and 11 -ing word instances respectively. These account for 45% of the total 141 -ing words evaluated for these subcategories.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Prog_PRact	52	If you are redirecting a restore operation, note the following:	復元操作をリダイレクトしたら、次を注目して下さい:
Prog_PRpas	11	Consider what type of Windows computer is being protected , the available hardware, and the system BIOS when selecting the type of bootable media to create.	作成するためにブート可能なメディアの種類を選択した場合利用可能なハードウェアおよびシステム BIOS かどのような Windows のコンピュータを 保護 されている考慮して下さい。

Table 3.71: Subcategories of the category Progressives with incorrect examples

The active voice present tense revealed two main errors. Firstly, 28 instances were analysed as being stylistically poor, that is, grammatical structures that would not be natural in the target language. The translations pertaining to this error category fell into the group of continuous tenses within conditional clauses. These clauses were generated by the MT system by attaching *すれば* (su-re-ba) to the noun denoting the English -ing word. However, it is considered more natural to use *ている場合* (te-iru-baai) for these cases. Secondly, 15 instances showed the incorrect use of the auxiliary verb for completed actions. The -ing words in the continuous tense are required to add the progressive aspect to the tense. The tense is given by the auxiliary *to be* which, in the examples analysed, is the present simple. However, on 15 occasions, *した* (shi-ta) was attached to the translation of the -ing words in Japanese, meaning that the action happened in the past and was completed. In its place, *ている* (te-iru), the auxiliary verb for the present progressive, should have been used. Additional issues included instances where the gerund-participles were translated as participial adjectives, intransitivity/transitivity ambiguities and terminological inaccuracy problems.

Error types for Prog_PRact	Examples	
	English source	Japanese MT translation
style ----- FR – 53.8%	If you are restoring in separate jobs, you must restore the Index database last.	別のジョブで復元すれば、インデックスデータベースの最後を復元しなければなりません。
auxiliary of completion ----- FR – 28.8%	If you are restoring Exchange Server 5.5, do the following:	Exchange サーバー 5.5 を復元したら、次をして下さい:

Table 3.72: Error types found in the translation of the subcategory of active voice present tense

The 11 examples classified as incorrect for the present tense passive voice Progressives displayed several issues. 4 examples were stylistically incorrect, similarly to the active voice subcategory. The terminology used to translate *snap* was inaccurate, as it was translated as *stop* for 3 examples. A particle to denote an object was incorrectly used to denote the patient of the passive verb on 2 occasions. Finally, 2 examples showed dependency errors caused by coordinative clauses.

Error types for Prog_PRpas	Examples	
	English source	Japanese MT translation
style ----- FR – 36.4%	If the restore is being redirected , see "Redirecting restores for SQL" on page 1356.	復元がリダイレクトされたら、「SQL のための復元をリダイレクトすること」を p.1356 の参照して下さい。
terminology ----- FR – 27.3%	If the cache file is located on a different volume than the volume that is being snapped , Backup Exec uses the following calculations:	キャッシュファイルが 止められている ボリュームより異なるボリュームで見つければ、Backup Exec は次の計算を使います:
particle ----- FR – 18.2%	Consider what type of Windows computer is being protected , the available hardware, and the system BIOS when selecting the type of bootable media to create.	作成するためにブート可能なメディアの種類を選択した場合利用可能なハードウェアおよびシステム BIOS かどのような Windows のコンピュータを 保護されている 考慮して下さい。
coordination ----- FR – 18.2%	Use the Exchange System Manager utility to manually dismount any databases that are being restored or check Dismount database before restore when creating the restore job.	手動で 復元されるか 、または点検が復元ジョブを作成した場合復元の前にデータベースをマウント解除するデータベースをマウント解除するために Exchange の System Manager ユーティリティを使って下さい。

Table 3.73: Error types found in the translation of the subcategory of passive voice present tense

ADVERBIALS

-ing words acting as heads of adverbial clauses introduced by prepositions or subordinate conjunctions were responsible for 51 incorrect machine translations, 12% of the total 414 evaluated. Adverbials of time introduced by *when* followed by an -ing word and adverbials of purpose introduced by *for* followed by an -ing word were the subcategories with the highest number of incorrect translations with 21 and 12 instances respectively.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Adv_T_when	21	When creating a script file, do not include all entries	スクリプトファイルを作成した場合、すべてのエントリを含まないで下さい
Adv_PU_for	12	Time limit used for determining No Comm communications status.	Comm コミュニケーション状態を 判断しない ために使われる制限時間。
Adv_M_by	5	Continue creating the job by following the procedures in "Creating a backup job by setting job properties" on page 361.	続行しまジョブを「バックアップジョブをジョブプロパティの設定によって作成します」の手順に p.361 の 続くこと によって作成します。
Adv_M_without	4	You can integrate Lotus Domino database backups with regular server backups without separately administering them or using dedicated hardware.	別にそれらを 管理するか 、または専用ハードウェアを使用しないで規則的なサーバーバックアップが付いている Lotus Domino データベースバックアップを統合できます。
Adv_T_after	3	Copies are made only after running non-AOFO (Advanced Open File Option) backups of the master and model databases.	複製はマスターおよびモデルデータベースの 連続した non-AOFO (Advanced Open File Option) バックアップの後やっとな撮られます。
Adv_T_before	2	The TSM server must be prepared using the BEX.MAC macro before becoming a TSM client.	TSM サーバーは TSM の顧客に 似合う前 の BEX.MAC のマクロを使用して準備されなければなりません。
Adv_T_while	2	Adding a media server to the list expands the number of media servers that can back up the Macintosh computer, while deleting a media server removes it from the list of media servers to which it advertises itself.	リストへメディアサーバーを追加することはメディアサーバーを 削除している間 からそれをそれ自身を広告するメディアサーバーのリスト削除する Macintosh のコンピュータのバックアップを作成することができるメディアサーバーの番号を展開します。
Adv_E_0	2	With True Image Restore, Backup Exec automatically restores data sets sequentially, beginning with the most recent set and going backward until the last full data set is restored.	TIR によって、Backup Exec は最新のセットおよび逆方向に移動 にはじまって 自動的に最も最新の完全なデータセットが復元されるまでデータセットを、順次復元します。

Table 3.74: Subcategories of the category Adverbials with incorrect examples

Similarly to the progressives category, the adverbials of time also displayed 12 instances with the incorrect use of the auxiliary verb for completion. した (shi-ta) was attached to the translation of the -ing words, meaning that the action happened in the past and was completed. Instead, する (su-ru) which indicates present time, should have been used for the translation of the structure *when* + -ing. Another 6 examples were deemed unintelligible by the linguist, as dependency errors made it difficult to understand the sentence. Additional issues arose with the use of particles, terminology and the translation of *when* as an interrogative pronoun.

Error types for Adv_T_when	Examples	
	English source	Japanese MT translation
auxiliary of completion ----- FR – 57.1%	When creating a script file, do not include all entries	スクリプトファイルを 作成した 場合、すべてのエントリを含まないで下さい
dependency ----- FR – 28.6%	The first decision to make when planning a production installation is to select the database server to use.	生産のインストールを 計画する ことが使うためにデータベースサーバーを選択することいつであるか作る最初の決定。

Table 3.75: : Error types found in the translation of the subcategory of when + ing

The adverbial of purpose introduced by *for* displayed a number of different translation errors. 3 examples showed dependency errors and 3 disambiguation problems with gerund-participles translated as participial adjectives. For 2 examples the preposition *for* was incorrectly translated and for another 2 the word class following the preposition was incorrect. Additional issues involved terminological inaccuracies.

Error types for Adv_P_for	Examples	
	English source	Japanese MT translation
dependency ----- FR – 25%	In the Clean-up options dialog box, review the settings for purging repaired and quarantined files.	クリーンアップオプションのダイアログボックスでは、修復されるページすることおよび検疫ファイル のための 設定を見直して下さい。
modifier ----- FR – 25%	"Checklist for troubleshooting devices that have gone offline" on page 1544	「オフラインになったトラブルシューティングの デバイスのための チェックリスト」 p.1544 の

Table 3.76: : Error types found in the translation of the subcategory of for + ing

CHARACTERISERS

-ing words functioning as modifiers were translated incorrectly in 48 examples, 9% of the total 536 evaluated. Pre-modifiers were the subcategory with the highest number of incorrect examples with 26, followed by reduced relative clauses with 15 and finally, 7 examples for nominal adjuncts.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Char_PR	26	There are three categories available for tracking media:	追跡のメディアのために利用可能な3つのカテゴリがあります:
Char_POrr	15	Insert the CD containing your clone CD image into the CD drive.	クローン CD イメージを含んでいる CD ドライブに CD を挿入して下さい。
Char_POnn	7	For more information on using SMS, see your Microsoft Systems Management Server documentation.	詳しくは SMS の使用で、Microsoft 社 Systems Management Server のマニュアルを参照して下さい。

Table 3.77: Subcategories of the category Characterisers with incorrect examples

The most prominent difficulty for the subcategory of pre-modifiers was ambiguity problems resulting in dependency issues with 13 examples classified as incorrect. The examples showed difficulty in allocating the correct head-modifier, subject-verb and subject-object dependencies. Terminological inaccuracy was responsible for 9 instances classified as incorrect. Additional errors included incorrect positioning of phrases and structures with stylistical issues.

Error types for Char_PR	Examples	
	English source	Japanese MT translation
dependency ----- FR – 50%	A SQLplus script in Backup Exec allows a default time-out of 10 minutes to handle the changing database state.	Backup Exec の SQLplus のスクリプトは 変更 のデータベースの州を処理することを 10 分のタイムアウトの既定値が可能にします。
terminology ----- FR – 34.6%	A media set consists of rules that specify append periods, overwrite protection periods, and vaulting periods.	メディアセットは規則から追記するピリオド、上書き禁止期間および アーチ形天井 のピリオドを指定する成っています。

Table 3.78: : Error types found in the translation of the subcategory of pre-modifiers

-ing words heads of reduced relative clauses were classified as incorrect in 15 occasions. 11 examples show dependency errors, where the relationship between

the -ing word and its near context was not correctly analysed. Of these, in 6 cases the -ing became the head of a noun phrase. In other cases the -ing word was translated as a modifier or using the progressive aspect tense. Additional errors emerged from the incorrect use of the auxiliary for completion and terminology.

Error types for Char_POrr	Examples	
	English source	Japanese MT translation
dependency ----- FR – 73.3%	Displays information detailing what has occurred while running the Command Line Applet and the specified option.	コマンドラインアプレットおよび指定オプションをか実行している間表示 情報の 起きた何が 詳述 。
dependency – ing as head ----- FR – 40%	Operating system currently running on this computer running the remote agent.	オペレーティングシステムの 遠隔エージェントを実行するこのコンピュータの現在 実行 。

Table 3.79: Error types found in the translation of the subcategory of reduced relative clauses

REFERENTIALS

The category of Referentials accounted for 40 incorrect examples, 30% of the -ing constituents evaluated for this category. The subcategory of catenatives showed 18 incorrectly translated -ing words, 12 belonged to the gerundial noun subcategory and 10 -ing words were functioning as objects of prepositional verbs.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Ref_cat	18	The Export message appears stating that the export has completed.	Export メッセージはエクスポートが完了したこと 表明 しますよう。
Ref_nn	12	Set time for automatic clearing of alert or disable automatic alert clearing	警告またはディスエイブルの自動警告の 清算 の自動清算の時間を設定して下さい
Ref_prepV	10	You can place a scheduled job on hold to prevent the job from running .	実行から ジョブを 防ぐ スケジュール済みジョブを保留にすることができます。

Table 3.80: Subcategories of the category Referentials with incorrect examples

-ing words functioning as objects of catenative verbs lose the relationship with the main verb. Two main translation patterns emerge. The -ing word is not translated as a catenated verb but as the nominal object of a verb, where the -ing word becomes either the head noun or a modifier (7). The main verb is translated as part of the previous noun phrase and the -ing word becomes the main predicate of the sentence (3

examples). Additional errors include the -ing word being translated as a reduced relative clause, intricate dependency errors and terminology inaccuracies.

Error types for Ref_cat	Examples	
	English source	Japanese MT translation
dependency - modifier ----- FR – 38.9%	Consider running full backups of mailboxes or public folders on a regular basis.	メールボックスまたはパブリックフォルダの 連続した 完全バックアップを定期的に 考慮して 下さい。
dependency - predicate ----- FR – 16.7%	Click Next to continue preparing disaster recover media.	クリック Next to は 続行 しま災害を回復しますメディアを 準備 します。

Table 3.81: Error types found in the translation of the subcategory of catenatives

The issue with the translation of gerundial nouns was the fact that the MT system did not analyse them as nouns. The analysis, therefore, was incorrect with gerundial nouns allocated as modifiers (2), gerunds (1), adverbials of mode (2), reduced relative clauses (1), verbs (1) or objects of incorrectly tagged nouns (1). Additionally, terminological inaccuracy was also a problem in 4 examples.

Error types for Ref_nn	Examples	
	English source	Japanese MT translation
modifier ----- FR – 8.3%	Automated restore selections and options checking , which tests the validity of your current SQL Server restore selections and job options before the restore job runs.	調べる 自動化された復元選択およびオプション復元ジョブ実行の前に現在の SQL Server の復元の選択およびジョブオプションの有効性をテストする。
gerund ----- FR- 16.7%	Click Sharing .	クリックの 共有 。

Table 3.82: Error types found in the translation of the subcategory of gerundial

The prepositional verbs in the sample are separated from their prepositional object by the first object. This causes the relation between the prepositional phrase and the verb to be lost. Therefore, the issues regarding the translation of the -ing words are several, such as their translation as modifiers, the separation of the sentence into unrelated chunks, or additional problems with coordinative phrases.

Error types for Ref_prepV	Examples	
	English source	Japanese MT translation
modifier ----- FR – 20 %	Select this option to prevent Backup Exec from overwriting files on the target disk with files that have the same names that are included in the restore job.	復元ジョブに含まれている同じ名前があるファイルが付いている作成先ディスクの 上書き ファイル から Backup Exec を 防ぐ ためにこのオプションを選択して下さい。
independent chunks ----- FR – 30%	The only way to prevent users in a profile from backing up a specific folder is to uncheck this option.	プロフィールのユーザーが特定のフォルダの バックアップ を 作成 することを このオプションのチェックマークをはずすべきである 防ぐ 唯一の方法は。

Table 3.83: Error types found in the translation of the subcategory of prepositional verbs

SUMMARY

The translation error types encountered within each group could be summed up as follows. The category of Titles revealed the same errors for all three subcategories. The -ing words were translated as adverbial clauses of mode introduced by the preposition *by*; the -ing words became subjects; and the -ing words were incorrectly analysed and translated as modifiers.

The category of Progressives was the worst-performing category in terms of the incorrect examples/frequency ratio. The main issue was the use of a stylistically poor structure, as well as the incorrect use of the auxiliary for completion.

Two subcategories were examined within the category of Adverbials: -ing words functioning as adverbials introduced by the conjunction *when* and introduced by the preposition *for*. Both subcategories revealed errors regarding the relation between the -ing word and its near context, which showed dependency errors. Additionally, the former included incorrect uses of auxiliary verbs for completion and the latter -ing words were incorrectly analysed and translated as modifiers.

Characterisers mainly displayed dependency errors whereby the -ing words functioning as modifiers were not analysed as such. As a result, the -ing words were translated as heads of clauses or were modifying incorrect heads. Terminological issues were also frequent.

Finally, the category of Referentials revealed ambiguity issues whereby gerund-participles were analysed as participial adjectives, and therefore, translated as modifiers. Also, in the case where the -ing words were analysed as gerund-participles, the correct function was not interpreted, and therefore, incorrect dependencies arose.

3.1.2.2 GERMAN

For German, 311 out of the 1,800 -ing words evaluated were classified as incorrect (17.28%). Similarly to Japanese, errors were spread across 20 subcategories. The following sections describe the performance for each category by observing the errors generated by the RBMT system.

TITLES

The category of titles included 557 examples, out of which 79 were classified as incorrect for German (14%). The worst performing subcategory is that of independent titles with 38 examples, followed by titles within quotation marks embedded in sentences or at the beginning of a sentence, with 23 and 15 instances respectively.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Title_ING_indp	38	Setting properties for Linux, UNIX, and Macintosh jobs	Einstellungs -Eigenschaften für Linux, UNIX und Macintosh-Aufträge
Title_ING_QM2	23	For more information, see " Customizing Connection Policies" on page 118.	Um weitere Informationen zu erhalten sehen Sie „Anpassen-Verbindungs-Richtlinien“ auf Seite 118.
Title_ING_QM1	15	" Viewing and changing active jobs" on page 452	„Anzeigende und ändernde in Arbeit befindliche Programmteile“ auf Seite 452
Title_ABING_indp	3	About troubleshooting DB2	Über Problemlösung DB2

Table 3.84: Subcategories of the category Titles with incorrect examples

38 examples were classified as incorrect for titles beginning with an -ing word. In 14 examples the gerund-participle was analysed and translated as a participial adjective. The -ing word was translated as a noun comprising a compound with the noun following the -ing word in 13 examples. Position errors appeared for 7 examples. The remaining incorrect instances were due to -ing words being translated as prepositions, stylistic issues and terminology inaccuracies.

Error types for Title_ING_indp	Examples	
	English source	Japanese MT translation
noun in compound ----- FR ⁴² – 34.2%	Formatting media in a drive	Formatierungs -Medien in einem Laufwerk
modifier ----- FR – 36.8%	Configuring and organizing columns in Backup Exec	Konfigurierende und organisierende Spalten in Backup Exec
position ----- FR – 21.9%	Searching for files to restore	Wiederherzustellen Suchen nach Dateien

Table 3.85: Error types found in the translation of the subcategory of -ing words at the beginning of independent titles

The titles starting with an -ing word within quotation marks embedded in a sentence displayed the same issues as the previous subcategory. 11 examples were translated as nouns comprising compounds with the noun following the -ing word, 6 were translated as modifiers and 5 were incorrectly positioned. The additional error was due to the -ing word being translated as a preposition.

Error types for Title_ING_QM2	Examples	
	English source	Japanese MT translation
noun in compound ----- FR – 47.8%	See " Checking Data Integrity" on page 996 for additional information.	Beachten Sie „ Überprüfungs -Datenintegrität“ auf Seite 996 für zusätzliche Information.
modifier ----- FR – 26.1%	For more information, see " Changing Windows security" on page 73.	Um weitere Informationen zu erhalten sehen Sie „ Ändernde Windows-Sicherheit“ auf Seite 73.
position ----- FR – 21.7%	See " Creating a synthetic backup by using the Policy Wizard" on page 876.	Beachten Sie „Ein synthetisches Backup mithilfe von erstellen der Richtlinien-Assistent“ auf Seite 876.

Table 3.86: Error types found in the translation of the subcategory of -ing words at the beginning of title within quotation marks embedded in sentences

Of the 15 examples of titles within quotations starting with an -ing word at the beginning of a sentence judged to be incorrect by the evaluators, 6 were translated as compound nouns, 6 were translated as modifiers and 3 were positioned incorrectly. In one example the -ing word was translated as a preposition. The translation options, therefore, are the same as the subcategories described above.

⁴² FR refers to the frequency rate of the translation errors.

Error types for Title_ING_QM1	Examples	
	English source	Japanese MT translation
noun in compound FR – 40%	" Setting catalog defaults" on page 495.	„ Einstellungskatalog -Standards“ auf Seite 495.
modifier FR – 40%	" Viewing and changing active jobs" on page 452	„ Anzeigende und ändernde in Arbeit befindliche Programmteile“ auf Seite 452
position FR – 20%	" Performing an automated restore by using the Disaster Recovery wizard" on page 1506	„Eine automatisierte Wiederherstellung mithilfe von durchführen der Disaster Recovery-Assistent“ auf Seite 1506

Table 3.87: Error types found in the translation of the subcategory of -ing words at the beginning of titles within quotation marks at the beginning of sentences

CHARACTERISERS

-ing words functioning as modifiers were translated incorrectly in 89 examples, 16% of the total 536 evaluated. Pre-modifiers were the subcategory with the highest number of incorrect examples with 51, followed by reduced relative clauses with 26 and finally, 12 examples for nominal adjuncts.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Char_PR	51	The alerting statements are Notes, Cautions, and Warnings.	Die alarmierenden Anweisungen sind Hinweise, Vorsicht und Warnungen.
Char_POrr	26	Insert the CD containing your clone CD image into the CD drive.	Fügen Sie das kadmiumhaltige Ihr CD Image der Duplizierung in das CD-Laufwerk ein.
Char_POnn	12	For more information on excluding files from Delta File Transfer, see "Configuring Global Exclude Filters" on page 1064.	Um weitere Informationen zu erhalten über ausschließlich der Dateien von der DeltaDateiübertragung, sehen Sie „Das Konfigurieren global schließen Filter aus“ auf Seite 1064.

Table 3.88: Subcategories of the category Characterisers with incorrect examples

The main difficulty for the subcategory of pre-modifiers was deciding whether the modifying -ing word should be translated as an adjective or within a compound noun. In 39 instances, the -ing word was translated as a separate modifier, when the combination of the -ing word with the modified noun was necessary to compose a compound noun. The remaining examples displayed instances where the -ing word was translated within a compound noun when it should not have been, or it was translated as an infinitive or a preposition.

Error types for Char_PR	Examples	
	English source	Japanese MT translation
separate modifier ----- FR – 76.5%	Depicts the starting slot.	Stellt den startenden Schacht bildlich dar.
compound noun ----- FR – 23.5%	If necessary, you can perform rolling upgrades in the CASO environment.	Bei Bedarf, können Sie Rollen- Upgrades in der CASO Umgebung durchführen.

Table 3.89: Error types found in the translation of the subcategory of pre-modifiers

26 examples of the reduced relative clause subcategory were classified as incorrect. Of these, 10 examples were translated modifying the noun following the -ing word and not the noun preceding it. 4 examples displayed positioning problems. Additional issues included the formation of compounds instead of separate modifying clauses, the use of incorrect tenses, and 3rd person singular verb and plural noun ambiguities.

Error types for Char_POrr	Examples	
	English source	Japanese MT translation
incorrect head modified ----- FR – 38.5%	To restore the cluster quorum to a node running Active Directory without using Backup Exec Cluster	Um das Clusterquorum zu einem Knoten wiederherzustellen Active Directory ohne Backup Exec zu verwenden ausführend bündeln Sie
position ----- FR – 15.4%	Specifies the slot numbers containing the media to be labeled.	Gibt die Schacht-Anzahlen an, welche die gekennzeichnet zu werden enthalten Medien.

Table 3.90: Error types found in the translation of the subcategory of reduced relative clauses

Of the 12 examples for nominal adjuncts classified as incorrect by evaluators, 5 -ing words were translated modifying the following noun instead of the preceding one. The remaining examples showed a number of issues, such as incorrectly translated prepositions and word classes for the -ing words.

Error types for Char_POnn	Examples	
	English source	Japanese MT translation
incorrect head modified ----- FR – 41.7%	See "Using Domain Groups to Manage DLO Permissions" on page 1001 for instructions on configuring DLO to use domain groups to manage DLO permissions.	Beachten Sie „Verwenden der Domäne-Gruppen, um DLO Rechte zu verwalten“ auf der Seite 1001 um Anweisungen zu erhalten über konfigurierendes DLO, zum der Domäne-Gruppen zu verwenden, um DLO Rechte zu verwalten.

Table 3.91: Error types found in the translation of the subcategory of nominal adjuncts

ADVERBIALS

-ing word heads of adverbials clauses introduced by prepositions or subordinate conjunctions were responsible for 78 incorrect machine translations for German, 19%

of the total 414 evaluated. Adverbials of mode introduced by the preposition *by* were the most problematic subcategory with 36 examples. Adverbials of time introduced by *when* followed by an -ing word and adverbial of purpose introduced by *for* followed by an -ing word, with 14 and 11 instances respectively, were the next most problematic subcategories.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Adv_M_by	36	Installing clients by using logon scripts	Anmeldeskripte der Clients mithilfe von installieren
Adv_T_when	14	This option is only available when performing full backups; otherwise, enter 0.	Diese Option ist nur verfügbar, wenn sie Gesamtsicherungen durchführt ; andernfalls eingeben Sie 0.
Adv_PU_for	11	In the Clean-up options dialog box, review the settings for purging repaired and quarantined files.	Im Reinigung-Options-Dialogfeld überprüfen Sie die Einstellungen für die reparierten und isolierten Dateien Bereinigungs .
Adv_M_without	6	Uses media auxiliary memory for inventory, which allows the media to be identified without having to be mounted.	Verwendet zusätzlichen Arbeitsspeicher der Medien für Bestand, der den ohne ermöglicht zu müssen identifiziert zu werden Medien, eingehangen zu werden.
Adv_T_before	5	The TSM server must be prepared using the BEX.MAC macro before becoming a TSM client.	Der TSM Server muss mit dem BEX.MAC Makro vorvorbereitet werden, bevor man ein TSM Client wird .
Adv_T_after	4	You must include a Sylink.xml file that gets created after installing and using ProductName Console.	Sie müssen eine Sylink.xml Datei enthalten, die erstellt erhält, nachdem es ProductName Konsole installiert hat und verwendet hat.
Adv_T_while	1	Number of errors encountered while trying to locate data.	Anzahl Fehler beim Versuchen angetroffen, Daten zu finden.
Adv_E_0	1	If you want managed media servers to be protected using a bootable tape image, you must run the IDR Preparation Wizard at each of the managed media servers where a bootable tape device is installed.	Wenn Sie die verwalteten mit möchten einem startfähigen Band-Image geschützt zu werden Medienserver, müssen Sie den IDR-Vorbereitungs-Assistent an jedem der verwalteten Medienserver ausführen, in denen ein startfähiges Bandgerät installiert wird.

Table 3.92: Subcategories of the category Adverbials with incorrect examples

Two main translation errors were observed for the structure *by + ing*. In 19 examples the translation of *by using* was the sequence *mithilfe von*, which was consistently placed before the verb and not before the object. Secondly, for 10 examples the use of pronouns was incorrect. The pronouns generated by the RBMT system either did not refer to anything or referred to the object instead of the subject. Additional errors were due to the incorrect use of prepositions or the generation of incorrect word classes for the -ing words.

Error types for Adv_M_by	Examples	
	English source	Japanese MT translation
position ----- FR – 52.8%	Users can be identified by using the following options:	Benutzern können mithilfe von identifiziert werden die folgenden Optionen:
pronoun ----- FR – 27.8%	By dynamically allocating devices as jobs are submitted, Backup Exec processes jobs quickly and efficiently.	Indem man dynamisch zuordnet , werden Geräte als Aufträge gesandt, Backup Exec-Prozess-Aufträge schnell und leistungsfähig.

Table 3.93: Error types found in the translation of the subcategory of *by + ing*

The 14 incorrect examples for the structure *when + ing* displayed a number of translation errors. 4 examples included an incorrect use of pronoun, which referred to the object instead of the subject. 4 examples were translated as modifiers, where on 3 occasions the translation of the -ing word modified the following noun and on one occasion the modifier had not modified the head. In 2 examples the translation of the -ing word was incorrectly located. A number of instances also showed an incorrect conjunction for the translation of *when*.

Error types for Adv_T_when	Examples	
	English source	Japanese MT translation
pronoun ----- FR – 28.6%	This option is only available when performing full backups; otherwise, enter 0.	Diese Option ist nur verfügbar, wenn sie Gesamtsicherungen durchführt ; andernfalls eingeben Sie 0.
modifier ----- FR – 28.6%	When searching for files to restore, or when viewing history logs, the DLO Administration Console accesses the network user data folders using the credentials of the currently logged in user.	Beim Suchen nach Dateien, um wiederherzustellen oder wenn , Verlauf anzeigend , protokolliert, die DLO Administrator-Konsole zugreift die Netzwerk-Benutzerdaten-Ordner mit den Identifikationsdaten des derzeit angemeldeten Benutzers.

Table 3.94: Error types found in the translation of the subcategory of *when + ing*

The 14 examples of the structure *for + -ing* classified as incorrect by evaluators showed three different translation errors. Firstly, in 5 examples the -ing word was

translated using a word class which did not agree with the preceding preposition. Secondly, in 4 examples the -ing word formed a compound noun with the following noun, where the -ing word was translated as the genitive of the noun compound it modified. Finally, in 2 examples the -ing word was translated as the modifier of the following noun.

Error types for Adv_P_for	Examples	
	English source	Japanese MT translation
word class ----- FR – 35.7%	Follow the instructions for the restoring Exchange data in "Requirements for restoring Exchange 2000, 2003, and 2007" on page 1248.	Befolgen Sie die Anweisungen für die wiederherstellenden Exchange-Daten im „Anforderungen für Exchange 2000 wiederherstellen 2003 und 2007“ auf Seite 1248.
compound noun ----- FR – 28.6%	Specify the path for a temporary staging location in the option titled Path on media server for staging temporary restore data when restoring individual items from tape.	Geben Sie den Pfad für einen temporären Inszenierung-Ort in der Option an, die Pfad auf Medienserver für temporäre Wiederherstellungs-Daten der Inszenierung betitelt wird, wenn Sie einzelne Objekte vom Band wiederherstellen.
modifier ----- FR – 14.3%	For more information on Minimum Rights required for querying data sources and privileges required for auditing, please refer to the Appendix A.	Um weitere Informationen zu erhalten über sprechen die minimalen Rechte, die für die abfragende Datenquellen und Rechte erfordert werden für Prüfung erfordert werden, bitte den Anhang A. an.

Table 3.95: Error types found in the translation of the subcategory of for + ing

PROGRESSIVES

Two subcategories from the Progressives category obtained incorrect judgments: the present active and passive voice subcategories, with 14 and 4 instances respectively. Of the 141 -ing words belonging to this category, therefore, 13% were classified as incorrect.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Prog_PRact	14	The volume containing the network user data folder is running low.	Das Laufwerk, das den Netzwerk-Benutzerdaten-Ordner enthält, ist niedrig laufend .
Prog_PRpas	4	If the restore is being redirected , see "Redirecting restores for SQL" on page 1356.	Wenn die Wiederherstellung sich umleitet , sehen Sie „Umleiten der Wiederherstellungen für SQL“ auf Seite 1356.

Table 3.96: Subcategories of the category Progressives with incorrect examples

The active voice present tense revealed a series of errors due to the progressive aspect tense not being recognised as such. In 3 examples the -ing word was translated as a modifier, twice modifying the noun following it and in one example not modifying anything. 3 examples displayed inflection issues. In another 3 examples the -ing word was translated into a noun. Additional errors involved incorrect positions, objects and agreements.

Error types for Prog_PRact	Examples	
	English source	Japanese MT translation
modifier ----- FR – 21.4%	There is a long-standing bug that is mis-converting and transposing hexadecimal and decimal numbers.	Es gibt einen althergebrachten Fehler, der ist , umstellend falsch konvertierend und die hexadezimalen und dezimalen Anzahlen.
inflection ----- FR – 21.4%	The Backup Exec client service, BEREMOTE.EXE, is not running on the computer	Der Backup Exec-Client-Dienst, BEREMOTE.EXE, führt nicht auf den Computer aus
noun ----- FR – 14.3%	If the current database server is not functioning , installing the new system as the database server is recommended.	Wenn der aktuelle Datenbank-Server nicht das Funktionieren ist , das neue System als der Datenbank-Server installierend wird empfohlen.

Table 3.97: Error types found in the translation of the subcategory of active voice present tense

REFERENTIALS

A total of 47 examples were classified as incorrect for the Referentials category, 35% of the total 134. The subcategory of catenatives showed 23 incorrectly translated -ing words; 14 belonged to the gerundial noun subcategory and 10 -ing words were functioning as objects of prepositional verbs.

Subcategory	№ of Incorrect Examples	Examples	
		English source	Japanese MT translation
Ref_cat	23	Click Next to continue preparing disaster recover media.	Klicken Sie auf "Weiter", um fortzufahren, Systemausfall vorvorbereitend , stellen Sie Medien wiederher.
Ref_nn	14	Tape Alert Warning	Nehmen Sie Alarm Warnung auf Band auf
Ref_prepV	10	A suite of pre-defined queries assist you in identifying key issues such as database integrity, security, and permissions tracking.	Eine Suite der vorbestimmten Abfragen unterstützen Sie, wenn sie Schlüsselfragen wie Datenbank-Integrität, Sicherheit und Rechnachvollziehen identifiziert .

Table 3.98: Subcategories of the category Referentials with incorrect examples

23 examples were considered incorrect for catenative -ing words. In 10 examples the verb into which the -ing word was translated was incorrectly placed. In 8 examples the -ing word was incorrectly translated into a modifier, either modifying the following noun or nothing at all. Additional errors included the translation of the -ing word into nouns and verbs, and issues whereby 3rd person singular verbs were translated as nouns, or past participles functioning as modifiers were translated as past tense verbs.

Error types for Ref_cat	Examples	
	English source	Japanese MT translation
position ----- FR – 43.5%	You can begin using Backup Exec.	Sie können Backup Exec, zu verwenden anfangen .
modifier ----- FR – 34.8%	Consider running full backups of mailboxes or public folders on a regular basis.	Erwgen Sie laufende Gesamtsicherungen der Mailboxen oder der allgemeinen Ordner regelmäßig.

Table 3.99: Error types found in the translation of the subcategory of catenatives

Within the gerundial nouns, 5 examples were translated as modifiers, 4 included incorrect terminology although the -ing word was translated into a noun, and 3 were translated into infinitives instead of nouns. Additional issues involved compounding errors.

Error types for Ref_nn	Examples	
	English source	Japanese MT translation
modifier ----- FR – 35.7%	Automated restore selections and options checking , which tests the validity of your current SQL Server restore selections and job options before the restore job runs.	Automatisierte aktivierende Wiederherstellungs-Auswahlen und -Optionen, das die Gültigkeit Ihrer Strom SQL-Serverwiederherstellungs-Auswahlen und -Auftragsoptionen prüft, bevor der Wiederherstellungsauftrag ausführt.
incorrect noun ----- FR – 28.6%	Schedule, CPS backup job running	Zeitplan, CPS-Sicherungsauftrag betrieb
infinitive ----- FR – 21.4%	Error messages are displayed in red and warning messages in orange to facilitate troubleshooting .	Fehlermeldungen werden im Rot und in den Warnmeldungen in der Orange angezeigt, um zu beheben zu erleichtern.

Table 3.100: Error types found in the translation of the subcategory of gerundial nouns

-ing words functioning as objects of prepositional verbs accounted for 10 incorrect examples. Of these, 6 showed that the verb into which the -ing word was translated was placed in an incorrect position. Other issues arose with the use of incorrect pronouns and tagging errors for the main verb being translated as a noun and the -ing word as verb.

Error types for Ref_prepV	Examples	
	English source	Japanese MT translation
position ----- FR – 60%	I would like to prevent files of specific types from being backed up.	Ich möchte Dateien der bestimmten Typen an sichert werden verhindern .

Table 3.101: Error types found in the translation of the subcategory of prepositional verbs

SUMMARY FOR THE GERMAN ANALYSIS

In summary, the translation error types encountered within each group are as follows. The category of Titles revealed that -ing words classified as incorrect were translated into nouns forming a compound with the noun following the -ing word or modifiers, or they were incorrectly placed within the sentence.

The main issue within the category of Characterisers was that some -ing words had to be translated into independent modifiers and some into compound nouns. This was particularly relevant for pre-modifying -ing words. The ambiguity between gerund-participles and participial adjectives was more pertinent for reduced relative clauses and nominal adjuncts. These also revealed dependency issues whereby the modifier was not modifying the correct head.

The three subcategories with the highest number of incorrect examples were analysed in the Adverbials category. The adverbial subcategory of mode introduced by the preposition *by* showed positioning and pronoun errors. Pronoun errors were also recurrent for the adverbial subcategory of time introduced by the conjunction *when*, apart from gerund-participle vs. participial adjective ambiguity issues. The adverbial subcategory of purpose introduced by the preposition *for* displayed errors related to the word class into which the -ing word was translated, compounding issues and, once again, ambiguity issues whereby gerund-participles were analysed and translated as modifiers.

Progressives revealed that the RBMT system had difficulty in recognising the progressive aspect tense. Instead, the -ing words were analysed as attributives and translated into modifiers or nouns. Inflection problems also emerged.

Finally, the category of Referentials showed similar issues. The subcategory of catenatives suffered from the incorrect analysis and translation of gerund-participles as participial adjectives. So did a number of gerundial nouns. Position was an issue relevant for both -ing words functioning as objects of catenative verbs and -ing words functioning as objects of prepositional verbs. Gerundial nouns, in turn, were badly affected by the use of inaccurate nouns.

3.1.2.3 SUMMARY FOR THE JAPANESE AND GERMAN ANALYSIS

Although the majority of -ing words for German and Japanese were grammatically and accurately machine translated by the RBMT system, 17% were still evaluated to be incorrect. The analysis of incorrect output revealed two main sources of errors. On the one hand, the difficulty of disambiguating gerund-participles from participial adjectives was observed. This, in turn, resulted in incorrect analysis and dependency errors. Moreover, further ambiguity exists within the gerund-participles. Being a broad category, as we described in Chapter 1 section 1.1, different functions are fulfilled by these -ing words and they must be differentiated in order to be correctly machine translated. This was not the case on many occasions and therefore dependency, word class, and location issues emerged.

On the other hand, we distinguished errors induced by the nature of the TL. In German, compounding is an efficient process for word formation. According to

Schiller (2005), 7% of the words of a newspaper text are formed through this process and it may even increase in technical manuals (e.g. 12% in a short printer manual). The option of using separate modifiers also exists. English noun phrases, therefore, can be translated into two structures. Often the RBMT system generates the incorrect structure and the output is incorrect. Pre-modifiers were affected by this issue. The positioning of the verb in the sentence is also a German-specific issue. German sentence structure is strict about the position of verbs. In a main clause, the conjugated verb must be the second element. In many subordinate clauses, on the other hand, it is placed at the end (Buck, 1999). This requirement often involves some reordering of the clauses in the sentence, which differs from the original English source, and poses difficulty for MT systems, as we saw with reduced relative clauses, catenatives and titles.

As for Japanese, issues regarding the information that is made explicit in the target language emerged. Although the underlying reason for these errors might be an incomplete or inadequate source analysis, the need to use particles to make subjects, main predicates, objects and the aspect of tenses explicit created problems for the RBMT system.

3.1.3 SUMMARY FOR THE HUMAN EVALUATION ANALYSIS

In general around 72% of -ing words are correctly handled by the RBMT system for German, Japanese and Spanish and just over half for French. The -ing categories showed varying degrees of correctness. However, a more in-depth analysis into the subcategories revealed that the errors were spread, with only a low number of subcategories consistently performing either well or poorly.

The analysis also showed two types of errors: grammatical and terminological. Grammatical errors, in turn, could be divided into two. On the one hand, we found -ing word level issues. These were related to the challenge of disambiguating between the three different -ing word types defined by Huddleston and Pullum (2002), as described in Chapter 1: gerundial nouns, participial adjectives and gerund-participles. This resulted in the incorrect analysis of the source text and eventually, the incorrect translation. On the other hand, constituent level issues were observed. We attribute the term constituent level issue to those cases where no -ing word issues appeared, that is, an -ing word functioning as a noun was not translated as a modifier, for instance, but

the output is still ungrammatical. Two main reasons might cause this. Firstly, a complex context where additional features known to be problematic for RBMT systems are coupled with -ing words, passive structures, for instance. Secondly, the RBMT system could successfully identify the subcategory to which the -ing word belongs, but transfer/generation issues apply.

Interestingly, terminology emerged as a problem. As described in Chapter 2, a customised version of SYSTRAN was used to obtain the machine translation output, which included customised UD's. Additionally, project-specific UD's were created using the standard terminology update process at Symantec. During the evaluation, evaluators were asked not to penalise the translation of an -ing word for terminological inaccuracies as long as the translation conveyed the same meaning as the source constituent. Regardless of the measures taken, however, terminological issues appeared across all subcategories. We will return to this in Chapter 6.

3.2 AUTOMATIC METRICS

We calculated NIST, METEOR, GTM, TER and character-based edit-distance for French, German, Japanese and Spanish (see Table 3.102 for results).^{43 44}

	TER	EditD	GTM	METEOR	NIST
Spanish	27.00	1.45	0.75	0.85	6.20
French	75.20	4.85	0.38	0.52	3.03
Japanese	38.13	1.02	0.80	N/A	5.22
German	44.26	2.31	0.60	0.39	5.81

Table 3.102: Overall automatic metric scores for each target language

⁴³ **NIST** calculated using the package available at

<http://www.nist.gov/speech/tests/mt/2008/scoring.html>

Command: perl mteval-v11b.pl -n -d2 -r <file> -s <file> -t <file> > <outfile>

METEOR calculated using the package available at <http://www.cs.cmu.edu/~alavie/METEOR/>

Command: perl meteor.pl -s SYSTEM -t <test.sgm> -r <ref.sgm> --modules "porter_stem"

GTM calculated using the package available at <http://nlp.cs.nyu.edu/GTM/>

Command: java gtm <MTfile> <REFfile> > <outfile>

TER calculated using the package available at <http://www.cs.umd.edu/~snoover/tercom/>

Command: java -jar ../tercom.7.2.jar -r <refFile.txt> -h <MTfile.txt> > <outFile.txt>

Edit-distance calculated using the NLTK package available at <http://www.nltk.org/download>
Applied through a script to automate the nltk.evaluate.edit_dist module for files.

⁴⁴ Note that the scores for NIST, TER and GTM for Japanese were calculated with tokenised output, whereas for character-based edit-distance non-tokenised output was used to prevent the white spaces introduced by the tokeniser from biasing the score. METEOR was not calculated for Japanese because no resources are available.

As can be seen, Japanese and Spanish generally obtained the best scores whereas French obtained the lowest. We report the Pearson r correlation between different automatic metrics for each target language.⁴⁵ This tests whether despite their different approaches to measuring quality, the metrics correlate on the degree of quality for each segment within each target language. The results showed moderate agreement on the overall quality of -ing words, ranging from 0.58 to 0.93 (see Tables 3.103-3.106). From this we can conclude that, regardless of the algorithm used to calculate the difference between the MT output and the reference translation, the overall score was moderately agreed upon. Therefore, the correlations between human and automatic metrics scores should be somewhat consistent across metrics.

Spanish	TER	EditD	GTM	METEOR	NIST
NIST	-0.72	-0.61	0.72	0.64	
METEOR	-0.80	-0.74	0.68		
GTM	-0.80	-0.65			
EditD	0.76				
TER					

Table 3.103: Pearson r correlation between automatic scores for Spanish

French	TER	EditD	GTM	METEOR	NIST
NIST	-0.76	-0.65	0.92	0.64	
METEOR	-0.66	-0.76	0.66		
GTM	-0.82	-0.67			
EditD	0.69				
TER					

Table 3.104: Pearson r correlation between automatic scores for French

German	TER	EditD	GTM	METEOR	NIST
NIST	-0.75	-0.58	0.79	0.78	
METEOR	-0.80	-0.69	0.82		
GTM	-0.93	-0.62			
EditD	0.60				
TER					

Table 3.105: Pearson r correlation between automatic scores for German

⁴⁵ See Appendix G for Spearman's rho coefficient and Kendall's tau coefficient correlations.

Japanese	TER	EditD	GTM	METEOR	NIST
NIST	-0.65	-0.68	0.77	N/A	
METEOR	N/A	N/A	N/A		
GTM	-0.83	-0.85			
EditD	0.76				
TER					

Table 3.106: Pearson r correlation between automatic scores for Japanese

Our main interest in obtaining automatic scores was to test whether they correlated with human judgements (H). Therefore, we calculated the Pearson r correlations between human judgements and each automatic metric (see Table 3.107).⁴⁶

	TER - H	EditD - H	GTM - H	METEOR - H	NIST - H
Spanish	-0.69	-0.57	0.65	0.56	0.51
French	-0.59	-0.51	0.60	0.53	0.56
Japanese	-0.53	-0.54	0.61	N/A	0.50
German	-0.48	-0.53	0.53	0.47	0.42

Table 3.107: Pearson r correlation scores between automatic metrics and target languages

The results showed moderate correlations between 0.51 to 0.69 for Spanish, 0.51 to 0.60 for French, 0.50 to 0.61 for Japanese and 0.42 to 0.53 for German. Correlations at sentence level reported within the MT community usually range from 0.2 to 0.6. Correlations between human judgement and TER are between 0.39 and 0.539 depending on the number of reference translations used (Snover et al. 2006). Correlations between human judgements and METEOR vary from 0.278 to 0.399 (Banerjee and Lavie 2005) and 0.493 to 0.55 depending on the number of reference translations for Snover et al. (2006). However, note that these measurements have English as target language. Callison-Burch et al. (2007), however, reported Spearman correlations calculated for target languages other than English. TER scored 0.589 for adequacy and 0.419 for fluency, whereas METEOR scored 0.490 for adequacy and 0.356 for fluency. Our results belong in the upper bound of the reported correlations, and therefore, suggest that the metrics show similar correlations, regardless of the evaluation unit (i.e. sentence vs. subsentential units).

Automatic metrics measure the similarity between two strings. When the strings are the same, automatic metrics assume that the quality of the MT output is the same as the reference, and therefore, assign them the highest score. Calculating when two strings

⁴⁶ See Appendix G for Spearman's rho and Kendall's tau correlation coefficients.

are equal is easy. Therefore, assigning top scores is easy. The challenge for automatic metrics relies on assigning the correct scores to strings that are different. Given the high number of correctly translated -ing words in our corpus, the following question arose: Does the high population of correct units distort the correlation between automatic metrics and human judgements? We will return to this issue in Chapter 6.

We noticed that some metrics correlated better with some TLs than others. For French, Japanese and German GTM was the metric which correlated better with the human judgements, with EditD also at the same level for German. For Spanish, TER was the best metric. The metrics with the lowest correlation scores were NIST for Spanish and Japanese, and character-based edit-distance for French, scoring 0.50 and 0.49, and 0.51 respectively.

We performed a further test to investigate whether automatic metrics could be used as an alternative to human evaluators to discriminate between “correct” and “incorrect” translations of -ing words. In order to examine that, we created different categories depending on the human scoring. We divided the examples depending on the number of “correct” votes each obtained. Having four evaluators, we obtained 5 categories where, in the worst case, none of the evaluators considered the example correct (0) and in the best case, all four evaluators considered the example correct (4). When one evaluator considered a translation to be correct, that corresponded to 1 on our x axis (See Figures 3.1-3.4); where two said it was correct, it was equal to 2 and so on. We then calculated the average automatic metrics score for each category.⁴⁷

⁴⁷ Given that the automatic scores express the results in different scales, we normalised them to be able to compare the trends. NIST, GTM and METEOR provide the results in a scale of 0-1, which was transferred to a % scale. TER and Edit-distance scores do not have an upper bound. For these metrics, the highest score, i.e. the worst-performing score, was taken as the upper bound and normalized to %. Note also that these latter two metrics score best when the result is zero, as opposed to NIST, GTM and METEOR, for which a zero score is the worst possible result. For a better visualization, the % scores were reversed to align with the NIST, GTM and METEOR trend.

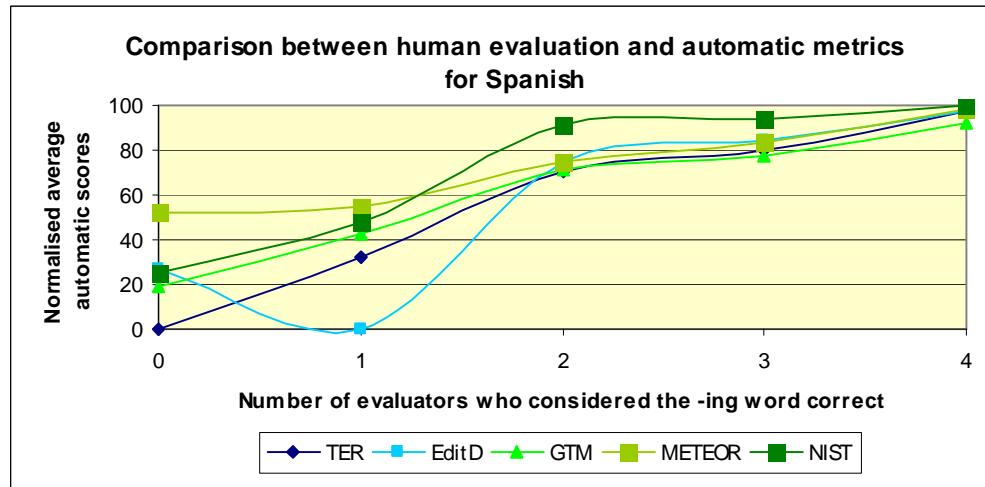


Figure 3.1: Comparison between normalised automatic scores averaged for the number of evaluators who considered the -ing words correct for Spanish

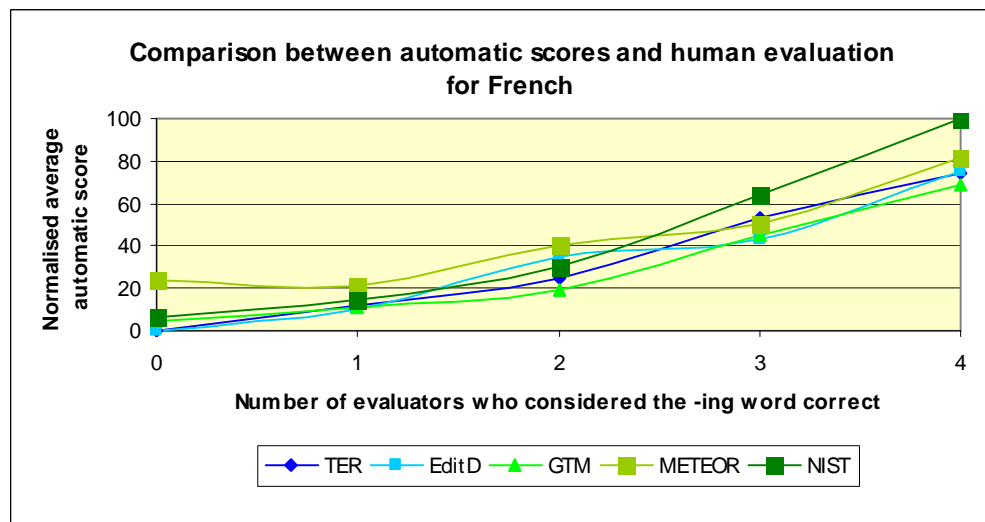


Figure 3.2: Comparison between normalised automatic scores averaged for the number of evaluators who considered the -ing words correct for French

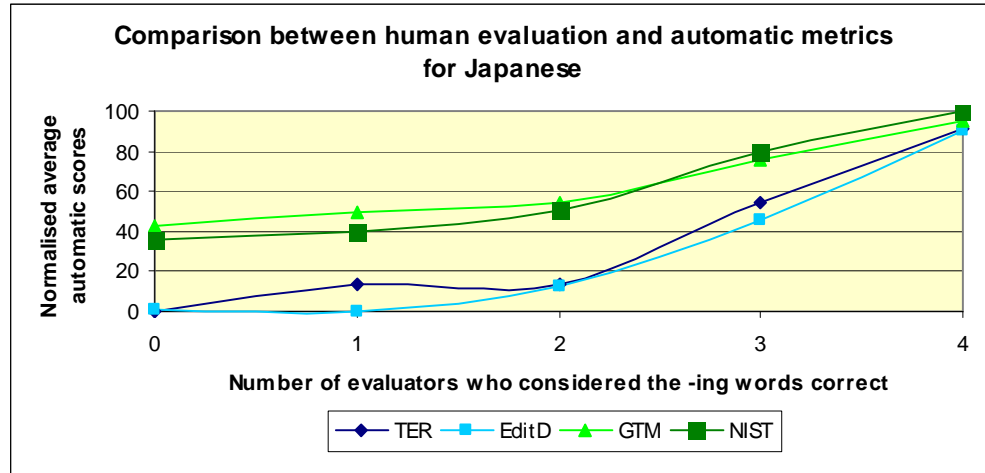


Figure 3.3: Comparison between normalised automatic scores averaged for the number of evaluators who considered the -ing words correct for Japanese

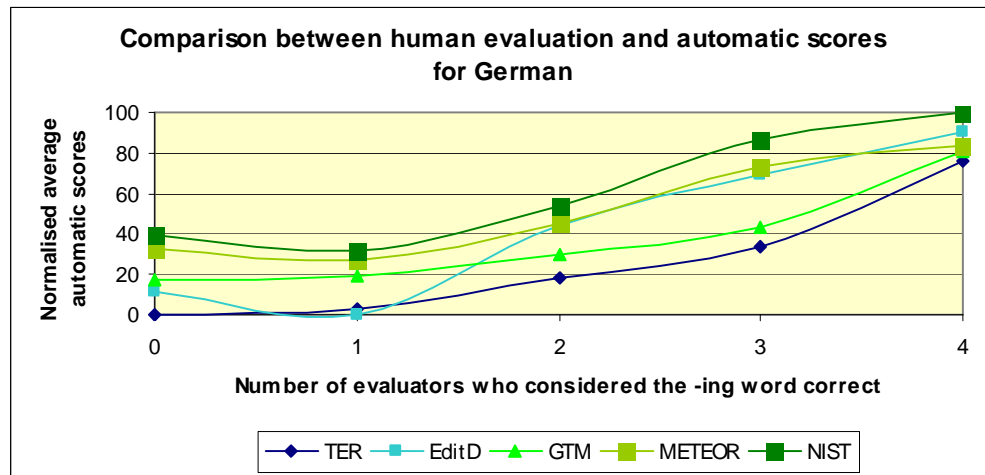


Figure 3.4: Comparison between normalised automatic scores averaged for the number of evaluators who considered the -ing words correct for German

We can see from the graphs that the tendency for each category correlated with the responses from the human evaluators. According to the automatic metrics, while the examples classified as incorrect by all the evaluators (0) needed a higher number of changes to convert to the reference translations, the examples classified as correct by all four evaluators (4) hardly needed any change. The only divergence in the trend

appears for edit-distance in category 1 for Spanish and German. A closer examination of the data showed that one evaluator rated the translation of the participial adjective "error-handling" as "de error-administración" as correct. This decision was probably made because the grammatical structure is correct - a *de* complementiser - although the terminology is incorrect. Due to the high frequency of this particular rendering (23 examples) and the high number of characters to be converted (from "de error-administración" to "de administración de errores"), the average trend for the examples in category 1 for Spanish was divergent from the trend. For German, position-related issues occur and they were not accounted for by the metrics at sub-sentential level.

We calculated the Pearson *r* correlations again based on the averages for each value on the x axis to verify the trend shown above (Table 3.108).⁴⁸ This showed that the agreement between human scores and the automatic metrics was high, varying from 0.86 to 0.98.

	TER - H	EditD - H	GTM - H	METEOR -H	NIST - H
Spanish	-0.97	-0.86	0.97	0.98	0.93
French	-0.98	-0.98	0.96	0.93	0.96
Japanese	-0.94	-0.92	0.96	N/A	0.96
German	-0.93	-0.94	0.92	0.94	0.97

Table 3.108: Pearson *r* correlation scores between automatic metrics and target languages calculated based on the averages obtained by grouping the examples according to the number of evaluators who considered them correct

The correlations showed that for French, TER and Edit-distance correlated slightly better with human judgements, with METEOR scoring lowest with a correlation of 0.93. This would suggest that a more direct calculation of the edit-distance, regardless of whether it is done on a word or character basis, correlates better with human scores than the measurement of precision, recall and their composite F measure or using stemmers. However, the results for Spanish displayed a preference for word-based metrics. METEOR was the metric which correlated better with the human judgements, closely followed by TER and GTM. Edit-distance was the worst performing automatic metric at 0.86. Japanese obtained the best correlation with NIST and GTM and German showed a slight preference for NIST.

⁴⁸ Note that the negative sign preceding the correlation scores for TER and character-based edit-distance means inverse association, not a poorer score. This is because higher values of human evaluation tend to be associated with lower values of these automatic metrics and vice versa.

The question that needs to be answered if we are to avail of automatic metrics as an alternative to human evaluation is: at what correlation level can we establish sufficient certainty to promote the use of automatic metrics rather than multiple human evaluations? We will return to this issue on Chapter 6.

3.2.1 SUMMARY FOR THE AUTOMATIC EVALUATION ANALYSIS

Automatic metric scores displayed a moderate correlation with human scores at a feature-level for all target languages. Furthermore, the automatic score averages for each of the categories obtained from the amount of correct votes assigned by humans revealed that automatic metrics seem to be able to distinguish between correct and incorrect -ing word translations.

3.3 CHAPTER SUMMARY

In this Chapter we learnt that around 72% of -ing words were correctly translated by SYSTRAN for German, Japanese and Spanish, and that just over half were correctly handled for French. In addition, the in-depth analysis allowed us to gather information about the correctly and incorrectly generated structures. We also established that there is reasonable correlation between different automatic metrics and our human evaluation, even on a subsentential level. But we noted that different TLs appeared to be more or less sensitive to specific metrics. This sets the path for selecting the subcategories of -ing words on which various improvement techniques will be tested. These techniques will be described in Chapter 4 and results will be presented in Chapter 5.

CHAPTER 4

CHAPTER 4: EXPLORATION OF APPROACHES FOR IMPROVING THE MACHINE TRANSLATION OUTPUT OF -ING WORDS

The aim of this Chapter is to explore approaches to improve the machine translation of specific -ing words which are applicable either before the text is submitted for translation or after it has been machine translated. The approaches are divided into two: Pre-Processing and Post-Processing. Under Pre-Processing, we investigate the use of Controlled Language rules (4.1.1) and Automatic Source Re-writing (4.1.2). Under Post-Processing we investigate Global Search & Replace of the target text (4.2.1) and Statistical Post-Editing (4.2.2). We describe the motivation for each approach and examine the strengths and weaknesses, as well as the procedures needed to implement them. Finally, we select specific -ing word subcategories and translation issues to report on the more in-depth requirements and details of each approach and test their performance. A human evaluation of the impact on machine translation quality of each approach is presented separately in Chapter 5.

4.1 PRE-PROCESSING

We mentioned in Chapter 1 section 1.5.1 that several methods have been devised to increase the machine translatability of texts. CLs appeared as one of the most widespread efforts, facilitating the improvement of source text quality for both human readers and machine translation. Alternative efforts also emerged to automate modifications in order to benefit MT systems, e.g. tagging. This section reviews the possibilities and requirements of CL rules to address a number of problematic -ing words identified during the human evaluation and explores the concept of Automatic Source Re-writing.

4.1.1 CONTROLLED LANGUAGE

The deployment of CLs in several industries was described in Chapter 1 section 1.5.1.1. In the following sections we focus on the rules constraining the use of -ing words. Suggestions for CL rules are presented considering results obtained during the human evaluation in Chapter 3. The process of rule creation is described as well as the performance measured.

As mentioned in Chapter 1, -ing words are often allocated a rule in CLs, most frequently banning the use of gerunds and present participles. In line with Kohl's (2008) claim, our human evaluation showed that not all -ing words were problematic. Certain -ing words were more problematic than others. In contrast to the general prohibitive approach, therefore, in this section we propose rules to address specific -ing word subcategories.

4.1.1.1 SELECTING -ING WORD SUBCATEGORIES FOR TESTING

Existing CLs often include a rule to prohibit the use of -ing words. In turn, Kohl offers a more flexible approach by reviewing ambiguous -ing words and offering possible alternatives. Certainly, the approach to take depends on the particular needs and expectations of each localisation workflow. In order to select a number of -ing subcategories to test the CL technique, we take the particular approach whereby CL rules are not used to address MT-specific weaknesses, but rather are more generic and could apply to other RBMT systems – see Chapter 1 section Current approaches to -ing words in CLs.

One of the main findings from the -ing evaluation was that a high number of examples were judged to be correct for German, Japanese and Spanish (reaching 72%), although French stayed at 52%. Based on these results, we argue that a broad approach of banning the use of -ing words in general taken by the majority of CLs is too restrictive. Firstly, we are now able to isolate the subcategories causing most problems. Secondly, the more targeted the rule, the more control over the performance, that is, the instances that are identified.

The first category in our classification were Titles. The evaluation showed that for French (61%), Japanese (32%) and Spanish (36%) titles were problematic. For German correct translations were more frequent, but still 20% remained incorrect. Two types of problems arose with titles. On the one hand, gerund-participles were analysed as participial adjectives and translated as modifiers into all target languages. On the other hand, generation problems were also observed. It would be possible to ban the use of titles starting with -ing words. However, this structure is so widely used in our corpus that such a rule would create confusion among writers and severe inconsistency with the existing documentation. Therefore, although it was acknowledged as an issue to be

addressed, we determined that CL was not the most appropriate way to tackle this category.

Characterisers were our second category. Whereas pre-modifiers were generally correctly analysed and translated in the cases where the terms were recorded in the RBMT system's dictionary, we observed that post-modifiers presented more problems. The MT output showed signs of correct post-modifying structures for the target languages. However, they were not successfully completed. For instance, for French (lowest correct percentage with only 63% correct output), passive voice reduced relative clauses were translated into a combined structure of passives and participles. For Japanese, post-modifiers tended to show dependency errors. Equally, the MT system often failed to create the correct prepositions and word classes following adjuncts across languages. The lack of terminological resources to generate appropriate modifiers in the target languages could be addressed by populating the user dictionaries. However, generating incorrect prepositions for specific nouns or word classes after prepositions was seen as a generation issue we could not address via CL. Finally, reduced relative clauses are structures that might be ambiguous for human readers and MT systems and that could be disambiguated by expanding them. We believed that CL was the correct solution for this later case.

The third group covered -ing words functioning as adverbials. The MT system's performance was average for these structures. Spanish and Japanese performed better with an average of 87% and 75% correct examples. German, in turn, obtained a lower number of correct examples, with a rate of 68%. However, the target language most affected by these structures was French, where only 56% of the examples were translated correctly. All target languages showed problems in the choice of preposition or subordinate conjunction. Japanese and German, in particular, also displayed ambiguity issues with gerund-participles translated as modifiers. We noted that even if evaluators were not particularly concerned about the ambiguity in the source sentences regarding implicit subjects, the output suffered from it. This was an issue that could be addressed via the CL. Making the subject of the subordinate clause explicit would reduce ambiguity and, presumably, increase machine-translatability. The use of incorrect subordinate conjunctions, which is language-specific, was a generation issue

that we could not address without further information on the internal behaviour of the MT system, and was therefore deemed to be unsuitable for CL.

The Progressives were our fourth category. For French and Spanish this group performed well with 74% and 82% of examples evaluated as correct. For German, the number of examples translated correctly was 72%, mainly due to inflection and agreement mistakes. For Japanese, the translation of these -ing words was mostly incorrect with only 40% of output correct. Most examples included stylistic issues, and could be described as understandable output but unnatural. Issues with dependencies also appeared, particularly with passive voice continuous tenses. French and Spanish, and to a lesser extent Japanese, showed problems when the progressive aspect was introduced in a passive structure. Seeing that passives posed problems across categories, we suggest that it might be an issue to be considered under an “avoid passives” CL rule, not an -ing-specific rule. Because the "avoid passives" rule should improve Spanish, French and Japanese performance and because Japanese was the only language seriously affected by this category, we considered that CL was not the most appropriate method to improve Japanese translatability.

Finally, let us review the group of referentials. This was by far the worst-performing group, with 61% correct examples for Japanese, 55% for Spanish, 47% for German and 40% for French. Despite the potential for improvement, however, we noticed that most issues were due to the lack of translation resources, e.g. terminology, recognition of specific patterns, etc. Gerundial nouns were incorrectly translated in cases where the MT system did not have the appropriate terminology available. The subcategory of catenatives included incorrect word classes for -ing words not because the system could not analyse the structures properly but because the translation equivalents in the target languages were sometimes unknown. Similar issues to gerundial nouns applied for phrasal verbs, whereas prepositional verbs behaved more like catenatives. However, apart from including word class and preposition choice problems, we observed that the particular structures within each subcategory performed differently for each target language. We opted for not banning the use of catenatives or particular prepositional verbs to compensate for MT weaknesses.

In summary, we proposed to address the following three issues via CL: ungrammatical use of determiners in the near context of -ing words, ungrammatical use of implicit subjects by using -ing words, and use of reduced relative clauses consisting of -ing words.

4.1.1.2 DESCRIPTION OF THE THREE -ING RULES

From the analysis of the evaluation results and the overview of the CL possibilities, we detected three different -ing word issues where CL rules could be beneficial. Therefore, we created three separate rules in acrolinx IQ to test whether machine translation could be improved.⁴⁹

The CL checker is based on pattern-matching and its rules can combine word-form and POS information, as well as additional extensive syntactic and morphological data obtained in the analysis process.⁵⁰ The rules are composed of three main sections: identification details, objects and patterns. A first section includes the necessary information to identify the rule and locate its help file. In a second section the author of the rule creates objects for each of the different items to be considered for the identification of a particular pattern. The final section contains the rules which represent the patterns we want the CL checker to identify (triggers), as well as exceptions (exclusion rules) (see Figure 4.1).

```
NAME identify_gerundial_nouns
RULE INFO
HELP LOCATION PATH

OBJECTS
ingl -> word with suffix ing
det -> article

RULES
TRIGGER: find ingl
EXCLUSION RULE: unless det ingl
```

Figure 4.1: Model of a simple rule for acrolinx IQ

CL-RULE 1: REDUCED RELATIVE CLAUSES

Reduced relative clauses are a grammatical structure in English which provide technical writers with the possibility of modifying a head in a condensed manner, e.g. *a file containing deleted data*, instead of *a file which contains deleted data*. However, they force the RBMT system to use disambiguation rules that would not be required if

⁴⁹ In consultation with Dr. Sabine Lehmann from acrolinx.

⁵⁰ acrolinx IQ suite uses the Penn TreeBank tagset.

a complete relative clause was used. Therefore, we propose suggesting to technical writers to expand reduced relative clauses into complete *that/which* clauses. To help in this task, the CL checker should be able to identify the cases where a noun is post-modified by an -ing word and suggest the use of a *that/which* clause.

As Kohl (2008) warned, “*noun + -ing word*” is not necessarily a reduced relative clause. This means that a rule which identified this general pattern would trigger a large amount of false positives. In order to avoid this, we decided to create a very specialised rule which would only focus on specific -ing words. We examined our corpus to gather information on the specific -ing words usually found in reduced relative clauses. Among others, *containing* or *running* were the most common in sentences such as “[t]he name of the computer running the remote agent” or “[t]he drive containing the Backup-to-disk folder is full”. Next, we created a rule which identified nouns followed by these -ing words only.

CL-RULE 2: IMPLICIT SUBJECTS

This rule aimed at identifying the cases where the subjects of the main and subordinate clauses were different. For the instances which were flagged, the introduction of the subject would be suggested. The focus was placed on the subordinate conjunctions *when*, *while*, *before* and *after*, as they were the most frequent conjunctions followed by -ing words.

When implicit subjects were used, it was not always easy to know whether the subjects were the same or not just by looking at the sentences. However, we noted that when an imperative was the main verb of the main clause, the subject of the adverbial subordinate clause was the second person singular, that is, *you*, the same as the main clause, 94% of the time.⁵¹ Similarly, it was also observed that when the main subject was *you* and the main verb included a modal auxiliary verb, the implicit subject was

⁵¹ In the evaluation corpus, we counted 83 sentences in which the subordinate conjunction was *before*, *after*, *when* and *while*, for which the implicit subject was *you*. Out of the 83, for 50 the verb in the main clause was in the imperative form and for 28 the subject of the main clause was *you* and the verb included a modal auxiliary. A verb in the imperative form was only used on 3 additional occasions for sentences where the main and subordinate subjects were the same but not *you* in a total of 149 sentences analysed.

always *you* (see examples in Table 4.1).⁵² Therefore, initially, we flagged all instances of *when*, *while*, *after* and *before* followed by an -ing word. Next, we used exclusion rules to discard the two contexts described above.

Conjunction	English source
when	Select Display filegroups when creating new backup jobs.
	When using this command, <u>you must also use</u> the -arobotic library switch.
after	After checking the check box, <u>type</u> the destination database name.
	<u>You must include</u> a Sylink.xml file that gets created after installing and using ProductName Console.
before	Before implementing the IDR option in a CASO environment, <u>review</u> the following:
	In addition, if the monitor change journal is enabled, <u>you must disable</u> it in the registry before beginning the Lotus Domino server recovery.
while	While using DLO, <u>you can suppress</u> dialogs by checking the Don't show me this message again check box.
	To ensure that you have the latest status and settings at any time while using the Desktop Agent, from the Tasks menu, <u>click</u> Refresh.

Table 4.1: Examples where *you* is the subject of both the main and the subordinate clause completed by imperatives and modal verbal phrases as the nuclei of the main predicate

Due to the difficulty in identifying the subordinate conjunction and the main clause, we were forced to limit our efforts to very specific cases. We used anchor points such as sentence beginning and end, commas, imperative verbs and modal auxiliaries. The rule focused mainly on discarding instances where the subject of the main clause was *you*. As a result, cases where the subjects of the main and subordinate clauses were the same but not *you* were also detected. These subjects were mainly the names of the products, e.g. "*After running the database snapshot job, Backup Exec creates history and job log information to indicate the job's status*". This is a grammatical structure and therefore, technical writers should not be forced to modify it. However, the solution of including product names in the rule was considered too lexical and very product-oriented and was discarded. As a result, we designed a rule that might pose extra effort for technical writers in that grammatical instances might be unnecessarily identified for modification. Nevertheless, should these instances be modified by the writers, it would not be detrimental for the MT system. From the analysis of the main human evaluation we learnt that although the problematic translations were mainly ungrammatical sentences, even grammatical implicit subjects generated incorrect

⁵² Based on a consideration of 149 instances in the evaluation corpus where the subordinate conjunctions *before*, *after*, *when* and *while* are followed by an -ing word.

output. Therefore, although we aimed at reducing the number of ungrammatical instances, we knew that the modification of instances containing product names would also be of benefit for the MT system, particularly for French.

CL-RULE 3: INSERTION OF ARTICLES

It is a characteristic of technical writing to drop articles (Bernstein, 1981; McCaskill, 1998). This makes it more difficult for RBMT systems to analyse the source correctly, especially when the words have homographs with different functions, e.g. “*Select this option to prevent Backup Exec from overwriting files on the target disk with files that have the same names that are included in the restore job.*” Articles are key disambiguation cues as they only allow for noun phrases to follow them. Therefore, we suggested that the technical writers introduce articles according to grammatical requirements, particularly in instances where the gerund-participles were directly followed by a noun phrase with no determiners or pronouns.

When creating CL-Rule 3, we realised that our possibilities were quite limited. The rule was supposed to identify the cases where the articles were missing. However, according to English grammar, indefinite plural nouns, generic/mass nouns and proper nouns should not be preceded by an article. Therefore, the rule would have to limit itself to suggesting the insertion of articles before singular definite and indefinite nouns and plural definite nouns and noun phrases.

Finally, this rule identified -ing words directly followed by a noun where this noun was either the head noun of a noun phrase or a noun modifying the head noun. Next, a series of exclusion rules restricted the nature of the noun or noun phrase following the -ing word to ensure that no proper nouns, mass nouns and plural indefinite nouns were identified. Nevertheless we did not find a way to distinguish plural indefinite nouns from plural definite nouns and therefore, these had to be left undetected.

4.1.1.3 MEASURING RULE PERFORMANCE

The next step was to test the rules' performance. Wojcik et al. (1990) and Adriaens and Macken (1995) report on the *technological evaluation* performed to test the Boeing Simplified English Checker (BSEC) and the Simplified English Grammar and Style Checker/Corrector (SECC) developed for Alcatel Bell respectively. In order to follow

a standardised methodology, we drew on their experience and calculated Precision and Recall, measurements borrowed from the field of Information Retrieval (IR).

Precision (P) measures “*the ratio of relevant to total material retrieved by a query or set of queries*” (Meadow et al. 2000: 322). In the context of CL, a high precision value means that only a reduced number of irrelevant instances were identified together with the relevant CL rule violations.

$$P = \frac{\text{relevant items retrieved}}{\text{retrieved items}}$$

Recall (R) is “*the ratio of the number of relevant records retrieved to the number of relevant records present in the file*” (ibid: 323). A high recall value means that a large proportion of relevant CL rule violations were retrieved. Recall shows the extent to which the instances of -ing words we aimed at identifying were actually retrieved.

$$R = \frac{\text{relevant items retrieved}}{\text{relevant items}}$$

A tendency to favour recall over precision seems to emerge from the results reported by Wojcik et al. (1990) and Adriaens and Macken, (1995). BSEC scores 79% precision and 89% recall, and SECC scores 87% precision and 93% recall. Based on the “*user appreciation of the system*”, Adriaens and Macken argue that precision is more important than recall (ibid). They claim that being prompted with irrelevant rule violations can be *misleading* and even *irritating*, which could lead to writers rejecting the use of the checker. However, they continue, non-retrieved violations are not that visible and therefore accepted, or even unnoticed, by the writers.

FINE-TUNING: TESTING FOR PRECISION AND RECALL

In order to perform the measurements, we returned to our -ing corpus. We divided the -ing corpus into two sets: the development set and test set. We used the sample judged during the human evaluation (1,800 -ing words) as the development set (develSet) and the remaining set (6,538 -ing words) as the test set (testSet). Bearing in mind all the corpus design factors considered in the compilation of the -ing corpus, we concluded that the number and variation of -ing words included in these sets was sufficient to design rules which would be as effective in unseen data.

Whereas measuring precision is relatively easy, measuring recall involves greater effort. Precision is calculated by examining the retrieved instances and by identifying irrelevant cases. However, in order to calculate recall, the set used to test the rules must be thoroughly analysed: all the relevant rule violations must be known.

By using the develSet, we could avail of the results from the human evaluation reported in Chapter 3 to obtain the required information to measure precision, and in particular, recall. We assigned the binary relevance judgment, *relevant* (R) - should be identified - or *non-relevant* (NR) - should not be identified - to each -ing word (see Table 4.2) for comparison against the -ing words retrieved by the rules. The information for the testSet was obtained by manually assigning R / NR to the instances of the particular -ing word subcategory relating to each rule.

	CL-RULE 1		CL-RULE 2		CL-RULE 3	
	R	NR	R	NR	R	NR
develSet	84	0	51	98	24	0
testSet	297	0	183	365	N/A	N/A

Table 4.2: Number of relevant and irrelevant examples in the rule development sets to measure precision and recall

It was considered essential that after developing and fine-tuning the rules in the develSet and the testSet the rules be tested on unseen data. This would ensure that the rules are not set-specific and can deal with IT procedural and descriptive texts in general. We were provided with a different user manual divided into 1,055 XML files containing 164,219 words. The files were checked in batch-mode to measure rule performance. Not having any insight into the new data, it was decided that only precision would be measured.

CL-RULE 1: EXPAND REDUCED RELATIVE CLAUSES

Test on develSet

A recall rate of 97.62% and a precision rate of 100% was calculated for Rule1. There were two reasons why a 100% recall was not achieved: incorrect tagging whereby plural nouns preceding the -ing word were analysed as a third person verb by the CL checker and a case where an -ing word preceded by the attribute of the verb was not detected. Given the high precision and recall results, we judged that no more effort was necessary.

After the primary filegroup is restored, select the rest of the filegroup backup sets containing the latest full and differential backups.
If the media in the drive is no overwritable or appendable, a message is displayed requesting that you insert overwritable media.

Table 4.3: Relevant examples of reduced relative clauses missed by the rule

Test on testSet

Recall for the testSet decreased to 78.45% whereas precision remained high at 96.4%. For recall, tagger issues accounted for 1.6% of the non-retrieved -ing words, and the rest were due to a number of -ing words not being listed as -ing words potentially constituting a reduced relative clause. Precision was affected by some of the listed -ing words appearing in structures other than reduced relative clauses.

Issue	Examples
Tagger issue	Number of mounts occurring since the last cleaning job.
	Direct a copy of the actual data streams being sent to media by a SQL database to a local directory for later use.
	Utility partitions being restored must belong to the same vendor.
-ing not listed	After restarting , you may see warnings about services failing to start.
	Increase this percentage is you receive an error message stating that the AOFO is out of disk space.
	Controls displaying a mnemonic (an underlined letter) can be selected regardless of focus by typing ALT and the underlined letter.

Table 4.4: Examples of relevant -ing words not retrieved by the rule

We decided to incorporate the additional -ing words occurring in reduced relative clauses as they would be beneficial for future unseen data. The risk of including these -ing words was that most of them only appeared on one occasion and we had no data to confirm that other grammatical structures did not occur with the same -ing word preceded by a noun. In order to check whether precision was affected by this decision, we calculated precision and recall again. As expected, recall increased to 90.57% and precision decreased to 80.4%. Precision was slightly affected by the tagger's lack of accuracy and ambiguous sentences, which contributed to 1.33% of errors each. However, the main problem was generated by false positives, accounting for 16.92% of the errors.

Issue	Examples
False positives	Configuring media servers for robotic library sharing
	During and after job processing , job log and job history information is generated for each job that is processed at the managed media server.
Tagger issue	Continue creating the job by following the procedures in “Creating a backup job by setting job properties” on page 361.
	Avoid running any processes or programs that would write excessive data to the drive being protected.
Ambiguous	Complete the backup job options following the procedures described in “Creating a backup job by setting job properties” on page 361.
	System being protected .

Table 4.5: Examples of non-relevant sentences identified by the rule

We observed that most of the non-relevant instances were prompted by the -ing words *concerning*, *creating*, *following*, *including*, *monitoring*, *processing* and *sharing*. As the amount of recall loss was lower than the irrelevance introduced by precision errors, we deleted these -ing words from the list and calculated the final precision and recall values for the rule. Precision increased to 98.22% and recall decreased to 85.86%.

Test on unseen data

CL-Rule 1, identified 88 rule violations in unseen data. From these, 83 were relevant and 5 were false positives, which resulted in a precision of 94.34%.

CL-RULE 2: MAKE UNGRAMMATICAL IMPLICIT SUBJECTS EXPLICIT

Test on develSet

CL-Rule 2 showed a recall rate of 86.8% and a precision rate of 80.33 %.

The queries execute in the UNIX user context after authenticating the native user credentials.
In order for DLO to function properly in a firewall environment, network file shares must be visible after establishing a remote connection such as VPN.
In addition, when restoring individual documents, the creation date and modification date properties do not restore.
The default behaviour when deleting a message from a mail archive may differ depending on the mail application.
The number of errors encountered while trying to locate data.
This will allow you to test the fibre connections before designating a new database server.

Table 4.6: Examples of dangling subjects retrieved by the rule

The examples not retrieved by the rule were due to structures listed in the exclusion rules also pertaining to relevant instances.

“Creating a restore job while reviewing media or devices” on page 297
When selecting databases to back up, the databases must be local to the Lotus Domino server.
Before redirecting the restore of a SharePoint Portal Server database, the SharePoint Portal Server software must be installed on the target server.

Table 4.7: Examples of dangling implicit subjects not retrieved by the rule

Precision showed some cases of grammatical implicit subjects retrieved, but no additional unexpected grammatical structures. We found 6 cases where both subjects were “you” and 6 cases where the subject could be either “you” or “Backup Exec” (we cannot confirm without more context). For the former case, we observed that the main verbs were imperatives, a case covered by the rule, but with tagging issues. The imperative *read* was tagged as a past tense verb and *direct* as an adjective. There was also a case where “you” was followed by a present simple tense verb. Due to the low number of occurrences, our rule does not focus on them. In the cases where the main subject was “Backup Exec”, we noticed that the implicit subject could either be “Backup Exec” or “you”, making the sentence ambiguous. We did not tune the rule to discard sentences where the main subject was “Backup Exec” or any other product name because the implicit subjects are ambiguous and because they have both grammatical and ungrammatical implicit subjects. Also, the rule would become product-specific. Therefore, these cases were deemed as acceptable false positives.

Issue	Examples
Tagger issue	Before starting backups for Exchange, read the following recommendations for configuring Exchange to make it easier to restore from backups:
	When restoring files, direct those files to the virtual server or the controlling node of the resource.
Backup Exec as main subject	After renaming the file, Backup Exec creates a new BEServer log file using the original log file name and then continues logging BEServer information to that original file name.
	Before reverting the database, Backup Exec deletes all existing database snapshots, including those created with SQL 2005, with the exception of the snapshot used for the revert.

Table 4.8: Examples which are not dangling subjects retrieved by the rule

Test on testSet

We tuned the rule to accept instances of imperative verbs after commas tagged as other parts-of-speech and tested the rule on the testSet. Recall increased to 94.1% and precision was at 80%. We noted three main possibilities for the low recall. First, the

rule did not include the possibility of an adverb appearing between the subordinate conjunction and the -ing word. Secondly, the rule did not differentiate imperatives from infinitives in the sense that in both instances the verb is tagged as a "VB". Finally, some tagging errors were noticed whereby plural nouns were tagged as verbs. Concerning precision, non-relevant cases retrieved by the rule were mainly implicit subject structures where the main subject and the subordinate subject were the same.

In an effort to increase recall but particularly precision, we took three steps. First, we included the possibility of an adverb appearing between the subordinate conjunction and the -ing word. Secondly, we differentiated imperatives from infinitives by specifying that a "to" could not appear in front of the potential imperative. Thirdly, we broadened the POS tag possibilities for "restores" so that it was not constantly considered a third person singular verb. The results were the same for precision and recall was slightly higher at 92%.

Test on unseen data

CL-Rule 2 retrieved 97 rule violations. From these, 73 were relevant and 23 were false positives. Within the false positives, however, 20 cases (20.63%) were implicit subjects where the subjects of the main and subordinate clauses were the same. As we mentioned earlier, these cases are grammatical. It could be argued that detecting them means increased effort for the technical writers, who have to review them, whether they modify them or not. Yet, should they decide to modify them, the MT output would not be adversely affected as the structure would become more explicit.

CL-RULE 3: INSERT ARTICLES

A rule to identify missing articles is reported within the top-10 rules for both BSEC and SECC. In the case of SECC, the rule stands out for its bad performance, only reaching 32% precision and 75% recall (Adriaens and Macken, 1995). The authors describe that this rule "*relies heavily on exhaustive and correct coding of the mass-count distinction*" (ibid: 130). In particular, disambiguation problems exist with nouns which can be either *mass* or *count* nouns depending on context. In order to address this, *mass-and-count* nouns were encoded depending on their frequency in the sets.

Test on develSet

Our -ing classification does not include a category where all the instances of missing articles in the -ing context are gathered. The omission of articles is not specific to structures where -ing words occur and was therefore not set as a category. Precision is computed on the number of retrieved instances, and does not require the knowledge of the total instances that should be retrieved. Recall, however, does. Therefore, we manually examined the development corpus and identified the total number of articles missing either between a gerund-participle or a gerundial noun and the following word, or before a participial adjective or a gerundial noun or a noun phrase including them. The best query in the develSet showed a recall rate of 85% and a precision rate of 83%.

No recover - Place database in loading state
Configuring notification in CASO
Using resource discovery to search for new resources

Table 4.9: Examples of missing articles in the context of -ing words retrieved by the rule

Plural nouns assigned a third person verb POS (VBZ) or participles ending in *-ed* not considered as modifiers were mainly responsible for lowering the recall (see Table 4.10).

Issue	Examples
Plural noun assigned a third person verb POS	Load balancing configuration causes servers to share the client communications load, and automatically implements failover of one of the servers crash.
A -ed adjective taken as participle	Unregistering UNIX agent-based target machine from the Information Server

Table 4.10: Examples of missing articles not retrieved by the rule

Precision, on the other hand, was lowered mainly due to the incorrect recognition of coordinated plural noun phrases as singular noun phrases. Cases were found where it was suggested that the determiner be added in front of definite gerundial nouns (see Table 4.11).

Issue	Examples
Plural noun phrase	"About redirecting Exchange storage group and database restores" on page 1260
Definite gerundial noun	This product is protected by copyright and distributed under licenses restricting copying , distribution, and decompilation.

Table 4.11: Examples which are not missing articles retrieved by the rule

Test on testSet

The rule was tuned to recognise coordinated noun phrases with a plural head. In addition, lexical specifications were added to avoid plural nouns such as *restores* or *causes* from always being tagged as verbs. This showed an increased precision rate of 88.17%.

Test on unseen data

CL-Rule 3 identified 105 rule violations. From these, 86 were relevant and 19 were false positives, which resulted in a precision of 82%.

4.1.1.4 SUMMARY FOR CONTROLLED LANGUAGE

This section presented the selection of three particular problematic -ing word issues to test controlled language rules. The creation of the rules was described, together with the fine-tuning process and measures for rule performance reported. We saw that CL is an interactive technique whereby a checker can be tuned to retrieve CL rule violations. The rules can be written to address particular structures. Rule performance is calculated by precision and recall.

No standard exists for sufficient precision and recall percentages and therefore it must be agreed upon for each context. A recall and precision of above 80% in the develSet and testSet and a precision of above 75% in unseen data was obtained for all the rules. Further lexicalised rules would possibly increase these percentages. Similarly, further compensation for tagger inaccuracies would also improve the rules' performance. Developers at acrolinx (direct communication) recommended not deploying rules with precision and recall numbers below 70%. Adriaens and Macken (1995) report an overall performance of above 90% precision and recall after fine-tuning on a *real-life test suite* (ibid: 126). The performance for each of their specific rules, however, varies from 65% to 100%. The reported values for our new rules were considered acceptable for an initial test on the effect they would have on translation quality.

4.1.2 AUTOMATIC SOURCE RE-WRITING

We mentioned in Chapter 1 that the idea of modifying the source to increase machine translatability has long been explored. CLs could be considered one of the first applications of this method where technical writers manually modify the text while creating the source for publication. Nevertheless, automatic modifications have received little attention apart from some normalisation efforts which try to standardise different spellings and contractions (see Chapter 1 section 1.5.1.2). In this section, therefore, we explore the possibilities of automatically modifying the source text in order to improve the machine translation of specific -ing word subcategories.

The benefits of this technique are several. Firstly, it is applied after the source text is completed, and therefore, the modified text is different from the original text for publication. This opens up the possibility of modifying the source text into an ungrammatical SL, should this prove efficient. It is only the RBMT system that will use this text, and therefore, it can be written in any way that favours it, whether it reads well or not according to the human reader; machine translatability is its only goal.

Secondly, the resulting text being independent of the original source, it is possible to create as many new source texts for MT as required. When applying CLs, for instance, writers work on the text destined for publication, as their goal is to improve text quality for human readability and/or compliance with company style guidelines, as well as to increase machine translatability. It makes no sense to have more than one source text for publication - either in terms of management or costs. Applying changes to the source text via techniques such as CL, therefore, make it a requirement that all the modifications are beneficial - or at least neutral - to all target languages. By having an automated process resulting in not-for-publication texts, it is possible to apply both shared and language-pair-dependent modifications. From a pool of possible transformations, specific relevant sets can be applied depending on the language pair into which the text will be machine translated.

Thirdly, we must not forget the novelty of applying the modifications automatically. Once the relevant modifications are identified, encoded and fine-tuned, it should take little time (all depending on the processing capacity) to apply them in a translation

workflow. This surely minimises the time-consuming and costly task of applying changes manually and one by one.

Automation, however, does not come without a price. If we are to automate all modifications, no human intervention is possible. This means that the modifications must be very precise, so as not to introduce additional translatability issues. The high precision required might compromise the potential advantage this technique offers.

4.1.2.1 SELECTING -ING WORD CATEGORIES FOR TESTING

For this technique to be successful two requirements must be met:

1. The modification to be performed on the source text must consistently result in correctly generated RBMT output.
2. No human interaction must be required.

Not having access to the core rules of the RBMT system, the only ways to come up with modification options are either by examining instances of the same structure classified as correct during the human evaluation or by trial and error. By observing how the MT system performs with specific structures, it is possible to change word order or to replace one word for another. This often means somehow altering the source string towards a more TL-like structure or introducing additional elements to make relations or information explicit. Yet, care must be taken not to overdo the changes as, after all, the MT system is designed to analyse grammatical SL.

We have considered the pros and cons of automatic pre-processing and the requirements for success. Let us now take a closer look at the search for modifications by focusing on some of the problematic -ing categories.

ASR-RULE 1: SPANISH TITLES

One of the most problematic categories for all languages are titles starting with -ing words. We decided to explore their pre-processing possibilities to try to improve their machine translatability.

-ing words at the beginning of titles are translated into Spanish as infinitives, nouns and gerunds. Evaluators judged nouns and infinitives as correct, whereas gerunds were not. SYSTRAN offers the option of translating imperatives into imperatives or infinitives into Spanish and French. Therefore, should the -ing words in titles be in the

imperative form, we could set SYSTRAN to generate infinitives consistently. This would generate not only correct titles for Spanish, but also improve the consistency. The modification to perform during the pre-processing stage is therefore to transform the -ing words into imperatives.

ASR-RULE 2: FRENCH, GERMAN, JAPANESE AND SPANISH TITLES

Titles for French could not be addressed in the same way as Spanish because infinitives are not correct in French. Evaluators penalised infinitives and were in favour of nouns instead. The correct translation pattern for French titles would be a noun followed by a *de* complementiser.⁵³ If we recreated this pattern in the source and transformed all -ing words in titles into nouns and add *of* after them, we would obtain the desired French translation (see Table 4.12). We therefore had to find a way of making the RBMT system generate nouns by modifying the source as little as possible to minimise the complexity of the transformation rule.

French-like pattern	New French MT	Original French MT
Installation of bv Control for Microsoft SQL Server	Installation de bv Control pour le serveur de Microsoft SQL	Bv Control installant pour le serveur de Microsoft SQL
Connection of the DLO on a different Backup Exec Media Server	Connexion du DLO sur différent Backup Exec Media Server	DLO se connectant sur différent Backup Exec Media Server

Table 4.12: Transformation: -ing into noun + of

Transforming -ing words into nouns, however, is not a process that can be generalised. For instance, *installing* becomes *installation*, just as *creating* becomes *creation* and *configuring* becomes *configuration* by deleting -ing and adding -ation. However, *using* becomes *use/usage*, *backing up* becomes *backup*, and *adding* becomes *addition*, to mention but a few transformations. If we were to encode this modification into a rule, we would require a specific rule for each transformation group, resulting in a lexicalised approach. In order to avoid this, we tried maintaining the -ing words intact and changing the words preceding or following them to try to deceive the RBMT system into analysing -ing words as nouns (see Table 4.13).

⁵³ Note that it is important to include the *de* complementiser. Otherwise, by transforming the -ing word into a noun we would be interfering with the following noun phrase.

In a first attempt, we added the preposition *of* after the -ing words. This isolates the -ing word from the rest of the string and creates a pseudo-nominal adjunct with the -ing words as heads. However, the RBMT system generated the same forms it generated when the *of* was not present.

In a second attempt, we added the article *The* at the beginning of the title. Only noun phrases are preceded by determiners and this could be a way to make the RBMT system analyse the element after a determiner as a noun. Still, the translation did not vary. The determiner was omitted by SYSTRAN.

In order to reinforce the idea that the -ing word should be analysed as a noun, we introduced a random adjective in front of the -ing word. We were aware that introducing this element would mean that the translation would include an unwanted element. However, should this element be easily recognisable, it would be easy to find and delete it from the target text. Once again, however, the translation did not vary and the determiner was omitted.

We finally tested whether using the imperative form instead of an -ing word would bring any positive change. Should it do so, we could reuse the rule used to transform the -ing words for Spanish titles. Unfortunately, no benefits were found.

New source	New MT
Installing of bv Control for Microsoft SQL Server	Installer de bv Control pour le serveur de Microsoft SQL
Connecting of the DLO on a different Backup Exec Media Server	Se connecter du DLO sur différent Backup Exec Media Server
The installing of bv Control for Microsoft SQL Server	Installer de bv Control pour le serveur de Microsoft SQL
The connecting of the DLO on a different Backup Exec Media Server	Se connecter du DLO sur différent Backup Exec Media Server
The quick installing of bv Control for Microsoft SQL Server	Installer rapide de bv Control pour le serveur de Microsoft SQL
The quick connecting of the DLO on a different Backup Exec Media Server	Se connecter rapide du DLO sur différent Backup Exec Media Server
The install of bv Control for Microsoft SQL Server	L'installer de bv Control pour le serveur de Microsoft SQL
The connect of the DLO on a different Backup Exec Media Server	Le connecter du DLO sur différent Backup Exec Media Server

Table 4.13: Examples of transformations

Modifying the immediate context of the -ing word did not deceive the RBMT system into translating -ing words as nouns. It was deemed necessary to transform the -ing words into nouns and use a lexicalised approach.

As mentioned above, titles were problematic for all languages. We therefore tested whether changing the -ing words into nouns followed by *of* would also benefit the other languages. Should this be the case, we could apply the same rule for a number of target languages, which increases the efficiency of the rule. Table 4.14 shows that the translations for German, Japanese and Spanish are also correct and consistent. This means that this transformation could be applied when translating to all four target languages.

New source	New German MT	Original German MT
Installation of bv Control for Microsoft SQL Server	Installation des BV- Steuerelements für Microsoft SQL Server	BV installierend , steuern Sie für Microsoft SQL Server
Connection of the DLO on a different Backup Exec Media Server	Verbindung des DLO auf einem anderen Backup Exec-Medienserver	Eine Verbindung herstellen zu DLO auf einem anderen Backup Exec-Medienserver
New source	New Japanese MT	Original Japanese MT
Installation of bv Control for Microsoft SQL Server	Microsoft SQL サーバーの ための bv Control のイン ストール	Microsoft SQL サーバーの ための bv Control のイン ストール
Connection of the DLO on a different Backup Exec Media Server	異なる Backup Exec メデ ィアサーバーの DLO の接 続	異なる Backup Exec のメ ディアサーバーの DLO へ の接続
New source	New Spanish MT	Original Spanish MT
Installation of bv Control for Microsoft SQL Server	Instalación del control de la BV para el servidor de Microsoft SQL	La BV de evaluación controla para Microsoft SQL Server
Connection of the DLO on a different Backup Exec Media Server	Conexión del DLO en diverso servidor de soportes Backup Exec	El conectarse al DLO en un diferente servidor de soportes de Backup Exec

Table 4.14: Transformation: -ing into noun + of

ASR-RULE 3: TIME ADVERBIAL CLAUSES INTRODUCED BY *WHEN* + *-ING* FOR FRENCH AND JAPANESE

When creating the CLRule 2 to address adverbial clauses with implicit subjects, we discovered that the implicit subjects of sentences with imperatives and modal

auxiliaries in their main clause were in the second singular person. This was therefore a pattern that could be easily identifiable.

These -ing word subcategories were problematic for French, German and Japanese. The French translation for this particular structure was a gerund. German showed different types of errors, such as incorrect use of pronouns, which might be due to the implicit subject, but also incorrect positioning of the verb or incorrect disambiguation of the -ing words as participial adjective. In turn, for Japanese, the main problem was that the auxiliary verb for completion was not appropriate. Spanish did not suffer from the implicit subjects as the use of impersonal structures was possible for this language.

We tested whether making the *you* subject explicit for the subordinate clauses would improve the machine translation of these -ing word subcategory for the different languages (see Table 4.15). We observed that for French and Japanese, the translation improved whereas for German it did not. We therefore decided to apply and test this rule for French and Japanese.

Source	Original MT	New source	New MT
When using this command, you must also use the -arobotic library switch.	En utilisant cette commande, vous devez également utiliser - le commutateur arobotic de bibliothèque.	When you use this command, you must also use the -a robotic library switch.	Quand vous utilisez cette commande, vous devez également utiliser - le commutateur arobotic de bibliothèque.
When creating a script file, do not include all entries	En créant un fichier script, n'incluez pas toutes les entrées	When you create a script file, do not include all entries	Quand vous créez un fichier script
Source	Original MT	New source	New MT
Also, when using ADBO note the following:	Auch wenn ADBO Hinweis das folgende verwendet wird:	Also, when you use ADBO note the following:	Auch wenn Sie ADBO Hinweis das folgende verwenden:
Click this to select the logon account to use when connecting to the media servers in the list.	Klicken Sie auf dieses, um das Login-Konto auszuwählen, um beim Eine Verbindung herstellen zu verwenden zu den Medienservern in der Liste.	Click this to select the logon account to use when you connect to the media servers in the list.	Klicken Sie auf dieses, um das Login-Konto auszuwählen, um zu verwenden, wann Sie zu den Medienservern in der Liste eine Verbindung herstellen.
Source	Original MT	New source	New MT
When creating a script file, do not include all entries	スクリプトファイルを作成した場合、すべてのエントリを含んではいけない	When you create a script file, do not include all entries	スクリプトファイルを作成するとき、すべてのエントリを含んではいけない

When contacting the Technical Support group, please have the following:	テクニカルサポートグループに連絡した場合、次を持ちなさい:	When you contact the Technical Support group, please have the following:	テクニカルサポートグループに連絡するとき、次を持ちなさい:
--	-------------------------------	---	-------------------------------

Table 4.15: Transformation for French, German and Japanese

ASR-RULE 4: PROGRESSIVES FOR JAPANESE

The progressive category performed poorly for Japanese with only 40% of the examples evaluated as correct. We therefore decided to try this approach to solve it. The main issue with progressives for Japanese was that the auxiliary verb for completion was not appropriate in the given contexts (see Table 4.16 for examples).

Source	Original Japanese MT	Post-edited version
You use this option since you are filtering members with a specific value of sysadmin.	sysadmin の特定の値のフィルタのメンバーであるのでこのオプションを使います。	sysadmin の特定の値でメンバーをフィルタしているので、このオプションを使います。
If the current database server is functioning , you should consider specifying the current database server when installing SAN Shared Storage Option to the new system.	現在のデータベースサーバーが機能すれば、SAN SSO を新しいシステムにインストールした場合現在のデータベースサーバーを指定することを考慮する必要があります。	現在のデータベースサーバーが機能している場合、SAN SSO を新しいシステムにインストールするとき現在のデータベースサーバーを指定することを検討する必要があります。

Table 4.16: Original data for Japanese

A first trial removed the progressive aspect and displayed the present simple tense instead (see Table 4.17). A second trial (see Table 4.18) removed the progressive aspect and introduced a number of desiderative structures with different levels of politeness. The Japanese machine translation output, however, did not improve nor was the meaning changed. Further attempts were not performed as expertise in the target language was crucial for a fast and effective trial and error process and the author did not have access to such expertise during this stage of the research. That is not to say that this method is not without promise, if tackled by somebody with the necessary target language expertise.

Source	Original MT	Post-edited version
You use this option since you filter members with a specific value of sysadmin.	sysadmin の特定の値を用いるメンバーをフィルタ処理するのでこのオプションを使う。	sysadmin の特定の値でメンバーをフィルタしているのので、このオプションを使います。
If the current database server functions , you should consider specifying the current database server when installing SAN Shared Storage Option to the new system.	現在のデータベースサーバーが機能すれば、SAN SSO を新しいシステムにインストールした場合現在のデータベースサーバーを指定することを考慮すべきである。	現在のデータベースサーバーが機能している場合、SAN SSO を新しいシステムにインストールするとき現在のデータベースサーバーを指定することを検討すべきです。

Table 4.17: transformation of progressive aspect into simple tense for Japanese

New source	New MT
If you want to install Backup Exec through Terminal Services and the installation media is on a shared drive (CD-ROM or network share) you must install using a UNC path.	インストールしたいと思えばターミナルサービスおよびインストールメディアによる Backup Exec は共有ドライブにあります (CD-ROM かネットワーク共有) あなたが UNC パスを使用してインストールしなければならない。
If you would like to install Backup Exec through Terminal Services and the installation media is on a shared drive (CD-ROM or network share) you must install using a UNC path.	インストールすることを望んだらターミナルサービスおよびインストールメディアによる Backup Exec は共有ドライブにあります (CD-ROM かネットワーク共有) あなたが UNC パスを使用してインストールしなければならない。
Should you want to install Backup Exec through Terminal Services and the installation media is on a shared drive (CD-ROM or network share) you must install using a UNC path.	あれば共有ドライブにターミナルサービスおよびインストールメディアによって Backup Exec をインストールしたいと思うためにも (CD-ROM かネットワーク共有) あなたが UNC パスを使用してインストールしなければならない。

Table 4.18: Transformation of progressive aspect into several desiderative structures

The lesson learnt from this attempt was that much care must be taken not to transform the sentences in such a way that the meaning is changed. While searching for a pattern that would produce a correct machine translation, it is easy to lose the sense of the source reference. It is essential to keep in mind the intended meaning of the original.

4.1.2.2 RESOURCES FOR IMPLEMENTING AUTOMATIC SOURCE RE-WRITING

Basically, what this technique requires is the identification of a problematic structure in the source text and its modification. Therefore, we need a procedure that will allow us

to search the source text, identify the exact structure to be modified and apply the changes. This, as Somers suggested (see Chapter 1 section 1.5.1.2), is a very similar task to “post-editing” the source. A helpful procedure he mentioned at the time was the use of global search and replace options available in text editors. Even nowadays, the most standard text editors such as MS Word provide little flexibility for S&R options. However, other independent editors such as UltraEdit or EditPro offer the possibility of using fully-fledged regular expressions such as the ones supported by Perl when writing search and replace rules.

Rules are written based on the patterns found in a representative corpus. Even when maximum closure is achieved (see Chapter 2 section 2.1.1.1), as new products and features are included in a company's documentation, lexical coverage might not reach 100%. This would not pose any problem if the issues are created by a specific lexical word. But it might be the case that the problem is created by a structural pattern, regardless of the lexical items involved. Therefore, if the rules, which are manually written, are to be created in an efficient way and are expected to cover instances that fall within the patterns addressed even if unseen in the corpus, it is essential to achieve some degree of generalisability. Regular Expressions (Regex) offer a degree of generalisability. Although they operate at string-level, it is possible to define concepts such as “word” using white spaces as boundaries, or to cover an undefined number of characters until a particular character or character set is found. This allows the inclusion of morphological information such as suffixes or prefixes, for instance. Yet, by using Regex directly on the source text, we find ourselves trying to define grammar character by character.

This is not strictly a weakness of Regex, but of the information that is available when applying them. Should the source text contain extra information, such as POS or phrasal relation, the Regex could be written to take this into account. A combination of deeper grammatical knowledge and surface-level lexis would greatly increase the capacity of Regex to address this type of problem.

In the localisation workflow in place at Symantec there already exists a tool which performs a thorough analysis of the source text: the CL checker (see section 4.1.2). acrolinx IQTM could provide us with additional information on POS and morphology. Moreover, using the checker for automatic implementation of modifications would be

beneficial in that it could complement and build on the modifications performed by the writers. In a separate module, existing CL rules could be extended to automatically introduce modifications which do not need human disambiguation, or even to introduce ungrammatical modifications which could not be published but would suit the RBMT system.

However, once the relevant words have been searched for, they must be transformed or regenerated. As a string-based approach, Regex require a character-based transformation, that is, it must be specified which of the characters searched for should remain, be replaced, added or deleted. This means that for a particular pattern, there will have to be as many rules as the number of possible combinations of searches and modifications required.

The CL checker performs an informative analysis. Nevertheless, its transformation capacity is minimal. Although it is able to deconstruct a word to obtain information from it when analysing, the minimum unit for re-writing is the word. It is not possible, for instance, to maintain the lemma and add a particular suffix. It has no generation capacity that goes further than the recombination of identified objects or introduction of literal strings and it does not even support the use of Regex. Therefore, a rule must be created for each specific lexical combination which can occur in the pattern addressed.

A number of other linguistically-informed tools exist that could help increase the generalisability and incorporate linguistic information into this technique. Such are lemmatisers, taggers, parsers, morphological analysers or linguistic generators. For the purpose of setting up the test for machine translation performance of an automatic pre-processing technique, we used Regex and a CL checker.

If we consider the selected -ing subcategories for testing this approach, titles starting with -ing words are easily identifiable structures, whereas implicit subject temporal subordinate clauses introduced by *when* followed by -ing words are not. Titles could be described as strings starting with capitalised words ending in -ing with no end-of-sentence punctuation mark. This is simple enough to build into a Regex rule and we therefore decided to use this option.

For the particular cases of the *when + ing* structure, however, the search function would need to identify modal auxiliaries – which can be done listing all the modals – as well as imperative forms. For the latter, however, each verb would have to be listed. Apart from this, as we saw when writing CL-Rule 2, the possible locations in which the structure can appear in the sentence are various and complex rules must be written to distinguish them from other structures. When writing CL-Rule 2, we identified all implicit subordinate clauses starting with *before*, *after*, *when* and *while* followed by an -ing word, except when the subject of the main and subordinate clause was *you*. These exceptions are written in the exclusion rules. Therefore, CL-Rule 2 already contains rules which identify the cases we want to modify automatically. We only needed to reverse the exclusion rules - which unflag structures - into triggers - which flag structures. We opted for this approach to test this structure.

CREATING THE RE-WRITING RULES

Following the same methodology as per the creation of CL rules, we used the *develSet* as the development set. We then used the *testSet* to test and improve the precision and recall of the rules. Finally, the rules were tested on unseen data.

ASR-Rule 1: Spanish Titles

This rule must (1) identify titles, (2) identify the -ing at the beginning of title, (3) transform the -ing word into the imperative form. Titles can appear as isolated strings or embedded in sentences within quotation marks. Also, some titles start with *About + -ing*. Within UltraEdit, we therefore specified that we were looking for strings that had no end-of-sentence punctuation mark or were within quotation marks and whose first word ended in -ing with an optional *About* preceding it. However, as we already pointed out, not all -ing words go through the same transformation to become imperatives and so an approach specifying lexical items was required. By discarding -ing words that appear as participial adjectives more often than as gerundial-nouns, we listed 94 different -ing words. These were divided into 8 groups depending on the modifications they had to go through to become imperative forms.

Next, the rule was tested for precision and recall in the *develSet*. From the -ing words classified in the Titles category, 5 are participial adjectives, 14 have an end-of-sentence punctuation mark and 2 start with symbols, and therefore, were

discarded. In total, there were 536 relevant instances. The rule found 524 which were all relevant, resulting in a 97.76% recall and 100% precision.

We then tested the rule in the testSet. There were 2,042 instances classified under the Titles category. However, some of these contained participial adjectives at the beginning of a title (18), others included end-of-sentence punctuation marks (45), started with symbols (15) or included corrupted characters (10) and were discarded. The relevant instances were reduced to 1,954. We checked for -ing word coverage of the rule. We noticed that whereas 94 -ing words were included in the rule, the testSet contained 164. Therefore, we included the missing ones. 1,913 relevant instances were transformed. Recall was high at 98% and precision reached 100%.

Finally, the rule was tested on new data. It modified 598 instances, from which 15 were incorrect transformations. This results in a precision of 97.5%.

ASR-Rule 2: Spanish, French, German and Japanese Titles

Similarly to the previous rule, this rule must identify titles and the -ing at the beginning of a title. Now, in contrast, the -ing words must be transformed into nouns. We listed all -ing words within the titles category in the develSet and testSet and grouped them according to the suffix to be attached to the roots in order to transform them into nouns. We created 32 groups depending of the transformation they had to go through to become nouns. By combining this information and that for finding titles, we created the new rule.

We next tested its performance by measuring precision and recall. The develSet contained 536 titles starting with gerund-participle -ing words. The rule transformed 524 examples, out of which 15 were incorrect transformations. Precision was 97.14% and recall 97.76%. The instances not found were either preceded by a symbol or had end-of-sentence punctuation marks.

We then tested the rule on the testSet. From the 1,954 relevant instances, the rule transformed 1,913, from which 8 were not correct as the title comprised of only one word and for 54, the *of* was preceding a preposition part of the adverbial constituent following the -ing word. Precision was 94.73% and recall 98%. The instances not found were due to additional spaces found before the beginning of titles.

We finally tested the rule in unseen data to make sure precision remained high. 45 out of the 598 -ing words transformed were incorrect, resulting in a precision of 92.5%.

When reviewing the performance of the rule, we noted that the fluency in English of some of the titles decreased. Whereas *Installation of a new programme* sounds natural, *Uninstallation of a new programme* is less natural. Whereas it is completely acceptable in this technique to create ungrammatical output, the behaviour of SYSTRAN with these words remains to be tested and clearly unknown words, such as *uninstallation*, will not be translated, unless included in the UDs.

ASR-Rule 3: French and Japanese -ING words functioning as Adverbials introduced by *When*

The second structure to be tested by automatic source re-writing is implicit subjects. Specifically, this rule must identify the instances where the main verb of a sentence is in the imperative form or uses modal verbs, and is modified by a subordinate clause of time introduced by *when* followed by an -ing word. As mentioned in section 4.1.2.2, the CL-Rule 2 identified these instances using 16 exclusion rules. We reused them and converted them into triggers. However, in order to account for the weak generation capacity of the checker, we had to create these 16 triggers for each of the -ing words we aimed at addressing. We included the 32 unique -ing words within the *when* + -ing subcategory in the develSet and the testSet.

Next, we tested the precision and recall of the rule in the develSet. There are 67 instances of *when* + *ing* in the develSet. From these, 36 have *you* as implicit subject with *you* as main subject. We tested the rule on the set of 36 and acrolinx IQTM transformed 32 sentences out of the 36 relevant instances, with recall at 88.89%. Not found instances were due to additional phrases within commas being placed between the main verb and the adverbial clause or due to the main verb being in the present simple form. All instances were relevant, preserving precision at 100%.

We then tested the testSet. acrolinx IQTM transformed 114 relevant instances of the 121 total present. This left recall at 94.21%. However, all identified examples were relevant, which meant that precision was 100%.

We finally checked the precision of the rule for unseen data. acrolinx IQTM transformed 31 instances. 28 were correct transformations for clear *you-you* cases. The additional three transformations were ambiguous, as it was not clear, out of context, who the subject performing the subordinate action was.

4.1.2.3 SUMMARY FOR AUTOMATIC SOURCE RE-WRITING

The automatic source re-writing has shown potential for becoming a viable solution for improving machine translation quality. Given that the modified text is not for publication, it offers the possibility of writing pseudo-English that best suits the RBMT system. Regex and the CL checker have proven workable for the implementation. Yet, although the high precision required for the technique was achieved, further research on generalisability, and in particular, on generation is called for in order to reduce the initial time-consuming rule crafting.

It is worth noting that because one of the strengths of this technique is its automation, and there is no possibility of human interaction, the precision of the re-writing must be high in order not to introduce additional complexities to the source. This, as we saw in section 4.1.2.2, is possible by reducing the scope of the rules.

4.2 POST-PROCESSING

Once the source text is machine translated, a decision must be made as to whether the text should be post-edited and this depends on its desired quality and purpose. Traditionally, as mentioned in section 1.5.2.1, this task has been carried out by post-editors. The task of editing machine translated text, however, is considered repetitive and time-consuming. This section reviews two techniques for improving the machine translation quality of -ing words by automatically modifying the target text. It focuses on Global Search & Replace and Statistical Post-editing, describing the creation of rules and their performance for the former and the training process for the latter.

4.2.1 GLOBAL SEARCH & REPLACE

The changes to be made in post-editing vary depending on the MT performance and can go from a complete restructuring of a sentence to the replacement of a preposition. Due to the deterministic character of RBMT systems, the same source structures are

consistently translated into specific target structures. Therefore, errors are repetitive. Whereas bad quality output requires a human post-editor to check the source text and review the translation, there are cases where the changes to be made are minimal and do not require complex disambiguation processes. Human interaction is not necessary to fix the latter cases. Global Search and Replace (S&R) modifications aim at automating these to eliminate them from the post-editors' tasks.

There are several benefits to applying this improvement technique at the post-processing stage. Firstly, anchor points from both target and source text can be used to write the modification rules. This allows mapping a particular error and modification to a particular source linguistic feature. Secondly, because the modifications are performed in the target text, no uncertainty as to the efficacy of the rule is introduced by the RBMT performance, as happened with the pre-processing techniques.

All errors can be corrected using a search and replace option on an individual basis, that is, it is possible to look for an incorrect string "*el función*" and replace it by the correct one "*la función*". However, the efficacy of this search/replace rule would depend on the frequency with which this particular string appears in the documentation. Further generalisability would increase the efficacy. Regex allow for a degree of generalisability, which means it can deal with unseen data. The use of regular expressions helps generalise the search and replace patterns. Regex take advantage of the characters, i.e. prefixes, suffixes, determiners, prepositions, etc. that remain constant in a particular structure and allow for different lemmas or words to be referenced.

4.2.1.1 SELECTING -ING WORD CATEGORIES FOR TESTING

Global S&R are applied in the target text, and therefore, it is based on the target errors that we must decide which modifications can be made without human intervention. From the analysis of errors performed in Chapter 3, we gathered the information about the types of errors specific -ing categories created. We observed errors at different depth levels.

Issues such as gerund-participles translated as modifiers would require a human post-editor to check on the -ing function in the source and modify the target if

necessary. This would usually require a complete revision of a clause or sentence, as the parsing of a gerund-participle as a participial adjective results in a completely different target tree to the original source meaning. The same occurs when a modifier modifies the wrong head or when clause/segment dependency errors appear. Nevertheless, more superficial issues were also observed, such as the generation of incorrect prepositions or articles in front of infinitives.

Let us describe the case of the generation of articles in front of infinitives, which reveals that the same target issue cannot always be addressed with the same modification, hence the need to map it to a particular source structure. 27 translations of -ing words for Spanish presented an article in front of a verb in the infinitive form. This is not grammatical in Spanish and therefore we considered it would be a good candidate where a superficial S&R would eliminate the problem. Guzmán (2008) described a Regex rule to address this error, which consisted of the deletion of the article. However, a closer look revealed that the nature of the article varied from -ing subcategory to subcategory. 15 out of the 27 examples belonged to the -ing category of Titles. The RBMT system generated infinitives preceded by articles, which were classified as incorrect by the evaluators. Given that infinitives without a preceding article were evaluated as correct, we could delete the article in order to obtain correctly translated -ing words. The remaining 12 words, however, belonged to different subcategories: 6 were objects of catenative verbs, 3 were gerundial nouns, 2 were reduced relative clauses, and 1 was a nominal adjunct. The Spanish catenative verbs required a noun object. The RBMT system did not generate nouns; instead, it generated infinitives preceded by articles. In these cases, therefore, the modification to apply would be the transformation of the infinitive into a noun. The gerundial nouns presented disambiguation issues resulting in incorrect analysis, which would need to be addressed by a human. The reduced relative clauses were translated into unintelligible ungrammatical structures, and so was the nominal adjunct. Therefore these instances would also need human intervention.

Based on the abovementioned analysis, we considered testing the technique through the titles translated as infinitives preceded by articles for Spanish (GSR-Rule 1) and objects of catenative verbs for Spanish and French, as both presented the same issue (GSR-Rule 2 and GSR-Rule 3 respectively). This would allow us to review the

conditions and benefits of using source and target text anchor points, as well as implementing improvements for a subcategory we did not yet address with other techniques.

4.2.1.2 RESOURCES FOR IMPLEMENTING GLOBAL S&R

In order to apply the S&R modifications we need aligned source and target texts. This can be obtained from TMs. Global S&R can be applied at the stage where all candidate translations are inserted in the TM and before this is sent to translators or post-editors. This is beneficial because the segments are aligned, as the technique requires, and because translators or post-editors would receive a text with fewer target language surface repetitive errors. Any environment that supports Regex can be used to write and test the rules on the TMs exported as .tmx files.

CREATING THE RULES

GSR_Rule 1: Articles in front of infinitives in titles

This rule aimed at eliminating the article in front infinitives at the beginning of titles for Spanish. We specified that for sentences without punctuation marks (titles) and starting with the capitalised article *El* followed by a word ending in *-ar*, *-er* or *-ir*, the article should be removed and the following word capitalised.

The rule was tested for precision and recall in the develSet. It identified all the 15 relevant examples, which resulted in a 100% precision and recall. Next, we calculated precision on the testSet. We restricted the performance measurement of this rule to precision because we had no information about the total relevant instances in the testSet. The rule identified 32 instances, which were all relevant. Similarly, the test in unseen data identified 6 instances which were all relevant. Precision remained at 100%.

GSR_Rule 2 & 3: Word classes following catenative verbs

GSR-Rule 2 and GSR-Rule 3 aimed at transforming the word classes of objects of catenative verbs. In order to do so, we first listed all the catenative verbs found in the catenative subcategory in our -ing corpus together with their corresponding Spanish and French translations. We then assigned each Spanish and French catenative verb the word class required for their objects. This would turn into the replacement pattern for the rule. Next, we listed the -ing words appearing as objects of catenative verb and their

Spanish and French translations in order to obtain the lemmas to be searched for and to which the modifications had to be applied. The observation of the type of incorrect output found for each catenative verb object was then carried out. Finally, we included the transformation rules required for each catenative verb object in combination with the particular lexical items.

The performance of the rules was tested next. We first tested the precision and recall of the GSR-Rule 2 and GSR-Rule 3 in the catenative subcategory of the develSet. There were 38 -ing words which occur as objects of catenative verbs. We transformed 20 instances for Spanish and 20 for French. All instances were relevant, and therefore, precision was 100%. Recall was lower, as not all objects could be searched for and modified with a simple Regex. Recall was 91% for Spanish and 83.33% for French.

We do not have information about the number of cases where the particular word sequences appear in the Spanish and French texts, and therefore, the number of relevant instances to be retrieved, necessary to calculate recall, was unknown. We restricted the performance measurement of these rules to precision. Yet, because we included all the catenative verbs and all the -ing words appearing as objects in our rules, we expect coverage to be wide. We acknowledge, however, that should the structures into which the -ing words were translated be different, they would not be identified by the rule. We tested precision for the testSet. Regex transformed 61 instances for Spanish and 78 for French. All examples were relevant, which meant that precision was 100%.

We finally checked the precision of the rule in unseen data. The rule found 14 instances for Spanish and 23 for French. The instances were all relevant.

4.2.1.3 SUMMARY FOR GLOBAL S&R

We showed that the application of Global S&R to automate superficial and repetitive post-editing through Regex is feasible. It provides control over the final output quality. We also learnt that it is necessary to map TL errors to source structures as the same error in the output can be due to different source structures, and therefore, require different modifications. This mapping relies on the consistent output of the MT system and therefore, should its output vary, the rules would no longer apply. This is particularly relevant in terms of terminology, as it can be easily modified using UD.

4.2.2 STATISTICAL POST-EDITING

As described in Chapter 1, SPE is based on the automatic acquisition of post-editing knowledge through the statistics obtained from a bilingual corpus of MT and their post-edited version. The automatic nature of this approach is an important strength, as it avoids the need for intensive textual analysis in order to create rules, as well as the effort required to craft the rules themselves, as with the previous approaches.

The SPE system is trained on an aligned parallel corpus of MT output and a corresponding reference sentence. The engine will try to improve new MT output based on the changes found in the training material. The quality of the corpus, therefore, is an important issue. The system can only make changes that have been derived statistically from the corpus. It is impossible for new types of modifications to be made. Therefore, for this approach a training corpus with a large coverage is essential.

SPE is a general approach to improvement. Whereas the previous approaches explored could be - and had to be - tailor-made to address a particular structure, SPE acquires the probabilities for modifications from the training corpus and no distinction is made regarding the type of modification. This has both advantages and disadvantages. On the one hand, being a general approach, there is no need to write complex rules every time a new structure/feature needs to be addressed. On the other hand, however, it is not possible to specify the type of modifications to be learned.

4.2.2.1 PREPARING THE SPE MODULES

The SMT system used to build the SPE module was Moses (Koehn et al. 2007).⁵⁴ Moses is an open-source factored phrase-based decoder for MT. By combining the decoder with a word alignment tool such as GIZA++ (Och and Ney, 2003a) and a language modelling tool such as SRILM (SRI Language Modeling Toolkit), it can be trained to translate between the desired language pair using any parallel corpus one might have (ibid). Being a phrase-based decoder, Moses maps contiguous word segments rather than stopping at single words. Being a factored decoder, Moses offers the possibility to customise the algorithm that learns the statistical probabilities so that it learns linguistically motivated factors, such as words annotated with POS or lemmas.

⁵⁴ www.statmt.org/moses

However, this requires having an annotated parallel corpus, which is not easily available for the sizes required in SMT, although the introduction of grammatical knowledge is a step towards a “hybrid” system, a more generalisable and deterministic engine, which would reduce the size of the training data required. We built the SPE engine following the “step-by-step guide” to Moses provided by the developers. This describes the baseline installation, training, tuning and translation steps. The factorisation and parameter tweaking possibilities, therefore, were not explored.

The first step for preparing an SPE module is to obtain and process the training corpora to use as input for the training modules. SPE only learns from seen data, and is not able, in its baseline use, to generalise from its learnings. Therefore, it is of the utmost importance to train it with texts as similar as possible to the texts that will be post-edited afterwards. However, the training material cannot be the same as the text that will be post-edited. In order to accommodate this, we used the TMs built from the latest version of one of the products included in the -ing corpus. We use sentences from our evaluation corpus to test the modifications SPE performs regarding -ing words. This corpus included texts from three different products but it mainly consisted of different documents for a particular product. The segments in the TMs, therefore, were similar to the segments we would be later post-editing. Additionally, we ensured that the SPE module was not trained with the same segments by discarding those included in the evaluation set.

The TMs for French, German and Spanish were easily accessible. However, Japanese TMs only accounted for a fifth of the product as compared to the European languages. In order to obtain a similar SPE engine, we would need to add content from other products to obtain a meaningful training corpus, which would not coincide with the main product included in the -ing corpus. Additionally, in an SPE experiment performed for Chinese, French, Italian and Japanese, Roturier (2009) pointed out that the results were not as positive for Japanese. One interesting observation made was that the translations in the MT were often somewhat “free”, i.e. not a strict translation of the source segment, which might have introduced noise in the training cycle. Given that the same training material could not be obtained for Japanese, that the training material obtained was too small to train an SPE module, and that, due to the TM management and localisation concept in Japan, the content in the TMs was not always

aligned and free translations were frequent, we opted to not perform the SPE experiment for Japanese.

Current research reports experiments on SPE engines built using different training corpora. Domain-specific corpora contain around 50,000 sentences or 500,000 words (Roturier and Senellart, 2008; Dugast et al. 2007; Díaz de Ilarraza et al. 2008; Isabelle, et al. 2007). General-domain corpora are larger, ranging from 700,000 sentences to 1.2 million (Roturier and Senellart, 2008; Dugast et al. 2007; Simard et al. 2007b). The results reported agree that domain-specific corpora considerably improve the RBMT performance using a relatively small amount of training data – as compared to SMT training data sets (Roturier and Senellart, 2008; Díaz de Ilarraza et al. 2008, Simard et al. 2007). We opted for the use of a domain-specific training corpus in the range of 50,000 sentences.

The TMs were made available in .tmx format, whereas what Moses requires is that the source segments (MT output) and the target segments (reviewed MT output), are separated into two sentence-aligned text files. The .tmx files included tags, both the ones inherent to the .tmx format and the ones inherited from the XML format in which the source text was written. .tmx format tags as well as XML tags identifying object types such as *guimenuitem* or *userinput* were deleted. Reference tags such as *ProductName* were replaced so that the sentences were complete. Finally, characters such as < or & were renamed so as not to conflict with other programming characters and XML format. Not doing so would have meant “losing” about 13% of the segment pairs, as over 7,000 contained tags. The final corpus contained around 56,500 segments (around 555,000 words) (see Table 4.19).

The TMs provided us with the reference translation – both post-edited and manually translated - of the target language. However, we still required the MT output. In order to obtain it, we machine translated the source segments in the TMs using a customised version of SYSTRAN. From the 56,500 set, we set aside 2,500 segments for the tuning cycle. After building the translation and language models, Moses carries out a tuning run in a different set of data based on minimum error rate training (MERT) (Och and Ney, 2003b) for BLEU. It uses the assigned probabilities and weights to translate the unseen set and fine-tunes them to obtain the best BLEU score. The use of automatic metrics such as BLEU for the optimisation process has been

challenged recently (Zaidan and Callison-Burch, 2009). Och and Ney (2003b) argued that the tuning should be performed using the same metric that will be used to evaluate the system's output after it has been built. Zaidan and Callison-Buch (2009) build on the idea by declaring that given that the aim of the system is to obtain human quality output and evaluation tasks involve a human evaluation part, the tuning should be performed using a human-based metric. They explore the possibility of using a parse-tree analysis of the translations to check against already-evaluated constituents to monitor their quality. Our SPE module might have benefited from the use of human evaluation scores during the optimisation cycle. Our experiment, however, predated this paper and its implementation, therefore, was out of scope.

After machine translating the source segments in the TMs into French, German and Spanish to obtain the RBMT output, and making sure that they were correctly aligned with the reference translations, we trained Moses following the step-by-step guide.. Using an Intel Pentium 4 CPU 3.20GHz and 2GB RAM, on Ubuntu 8.04.1, building the translation model with the 54,000 segment pairs took 40 minutes. The tuning time varied from 1 to 5 hours. Table 4.19 gives an overview of the training material for Moses.

	Spanish	French	German
Total number of segments in TMs	56,530	56,560	56,456
Number of segments in common with evaluation sample	27	33	27
Number of segments used for training after 50 word cut-off	53,778	53,737	53,931
Number of phrases generated in the SPE module	1,212,262	1,308,020	1,101,910
Number of segments used for tuning	2,500	2,500	2,500
Percentage of phrases filtered for "translation"	2.93%	3.33%	3.82%

Table 4.19: Training material for Moses

4.2.2.2 SELECTING AN -ING SAMPLE FOR TESTING

As opposed to the previous techniques, which were targeted at certain -ing word subcategories, we said that SPE constituted a general approach. The SPE module learns all changes to turn the MT output into its PE version. It would go against the principle of the technique and against its main strength to focus on specific changes. Also, it would not be fair to evaluate whether the SPE module performed changes on a particular -ing category, as the SPE module was not restricted to applying changes to -ing categories. Modifications could have been applied anywhere in the sentence. Therefore, we opted for not considering the -ing categories and for evaluating the

segments regardless of the -ing category to which they belonged. A re-evaluation of the whole evaluation corpus was out of the scope of this dissertation due to space, time and budget constraints. Therefore, we applied a simple random sampling (SRS) method to extract a representative set. According to Cochran (1963), 385 examples suffice to obtain a representative sample. We would therefore use this number to evaluate the impact on translation quality of SPE.

4.2.2.3 SUMMARY FOR STATISTICAL POST-EDITING

We trained a SMT system to automatically post-edit RBMT output. By training the system with the corpus size for domain-specific content suggested within the field, we aim at testing translation improvement in Chapter 5. SPE being a general technique, we do not aim at addressing specific -ing word subcategories. Instead, we will assess the improvement for -ing words in general, by evaluating a representative sample.

4.3 CHAPTER SUMMARY

This Chapter has reviewed four approaches for the improvement of machine translation output for -ing words. All of the approaches are independent of the RBMT system. Two of the approaches, controlled language and automatic source re-writing, pertain to the pre-processing stage, whereas global search and replace and statistical post-editing apply in the post-processing stage.

Applied during the pre-processing stage, CL and automatic re-writing work on modifying the source text, whereas the global S&R and SPE aim at improving the target output. CL, which is the only technique that can avail of human interaction, can only include shared problems across all target languages and is therefore restricted as a solution. Modifying the source to benefit one language only might not be cost-effective even when it is guaranteed that no degradations will occur for other target languages.

The advantage of automatic re-writing is that the resulting text is strictly for RBMT use and therefore no restriction is necessary on the number of modifications made for different target languages. It is also worth mentioning that because this new text is not for publication, in comparison with the other three techniques explored here, pseudo-language is acceptable as long as it benefits the RBMT system. The

disadvantage of this approach, however, was found to be the high precision required so as not to introduce additional complexities in the source.

Global S&R on the target output was also described as a technique. The benefits of this approach are that it works directly on the target text and can include anchor points from both the source and the target text to apply the rules. However, rules must be manually written for each target language.

An aspect that differentiates SPE from the other techniques is its general nature. Whereas for the other techniques specific rules must be written that address a particular structure, SPE does not target specific linguistic structures. Its main strength is that modification rules are learnt automatically and there is no need for a human to analyse and write the rules. However, we observed that because the probabilities are learnt from a corpus, its quality must be considered. Table 4.20 provides a summary of the characteristics of each technique studied.

	CL	Automatic	Global S&R	SPE
Modified language	SL	pseudo-SL	TL	TL
Human interaction	YES	NO	NO	NO
Source of error	shared	shared or TL-specific	TL-specific	TL-specific
End-product purpose	publication	MT	publication	publication
Scope	targeted	targeted	targeted	general
Resources for rule creation	SL	SL	SL + TL	raw TL + ref TL

Table 4.20: Summary of characteristics per improvement approach

Given the characteristics of each approach, we selected particular -ing structures to test each of the procedures. This Chapter presented the rule creation process and accuracy of rule performance by reporting precision and recall measurements. The analysis of the results for their effectiveness in terms of translation quality judged by human evaluators is presented in Chapter 5.

CHAPTER 5

CHAPTER 5: DATA ANALYSIS II

In Chapter 4 we explored techniques which could help improve machine translation. We further examined the potential benefits and conditions for each technique by applying them to a number of problematic -ing word subcategories. The rules for Controlled Language, Automatic Source Re-writing and Global Search & Replace modifications were crafted and the engine for Statistical Post-editing trained. Chapter 5 analyses the effect the techniques have on translation quality. In order to do so, this Chapter presents a human evaluation performed for each technique. The rules and engine described in Chapter 4 are applied and the translations evaluated. For each of the techniques we focus on the evaluation set-up, present the results and discuss the implications of deployment into a localisation workflow.

By applying the techniques, the majority of answers might fall into a single category. The Kappa inter-rater agreement scores might be skewed for such cases as Kappa scores are adversely affected by *prevalence* (Hunt, 1986; Di Eugenio, 2004; Viera and Garrett, 2005). In cases where the distributions of the answer categories are disparate, we expect high agreement, paradoxically, this situation results in lower Kappa scores (ibid). Therefore, we present the results of the human evaluation for the improvement techniques by reporting percentage agreements for the examples for which at least three of the evaluators agreed, as well as the Kappa coefficient scores.⁵⁵

5.1 CONTROLLED LANGUAGE

We wrote three CL rules – insert article, use subjects in subordinate clauses and expand reduced relative clauses – which worked with a precision and recall of over 80%. In order to measure their effect on the translation quality of -ing words, we applied them to the develSet. We configured the CL checker to detect the three new rules only and checked the develSet. acrolinx IQ™ identified 74 relevant examples for CL-Rule 1, 51 for CL-Rule 2, and 18 for CL-Rule 3. We re-wrote these sentences according to the guidelines provided to the technical writers in the help section. The irrelevant examples were discarded, as the technical writers would not modify them. A total of 143 -ing words were “controlled”.

⁵⁵ For comparison purposes and due to the type of evaluation performed, the results for the Controlled Language rules follow the same format used in the initial -ing human evaluation.

Note that these sentences included -ing constituents classified as correct, inconclusive and incorrect in the human evaluation in Chapter 3 (see Table 5.1). This is crucial for two reasons: firstly, to examine whether the examples previously judged to be incorrect or inconclusive benefit from the use of the new CL rules; and secondly, to make sure that the examples previously judged as correct are not adversely affected.

	CL-Rule 1 expand reduced relative clauses				CL-Rule 2 make implicit subject explicit				CL-Rule 3 insert article			
	FR	DE	JA	ES	FR	DE	JA	ES	FR	DE	JA	ES
correct	45	47	57	51	12	26	42	43	5	13	16	9
inconclusive	14	3	6	12	12	17	4	3	0	1	0	1
incorrect	15	24	11	11	27	8	5	5	13	4	2	8
total number of examples	74				51				18			

Table 5.1: Results obtained in the human evaluation for the examples addressed by the new CL rules

5.1.1 EVALUATION SET-UP

If we could reproduce the same evaluation setting in which the first -ing evaluation was performed, i.e. the same MT system version and resources, and the same evaluators, and if we could reproduce the same methodology, we would have the possibility of comparing results. The baseline quality of the RBMT output would be the same, the difference in quality of the new output would be due to the effect of CL rules. By hiring the same professional translators who participated in the first -ing evaluation we could provide continuity to the evaluation. The time gap from the initial evaluation to the second cycle (15 months) was long enough for the evaluators not to remember their judgements. However, the understanding of the evaluation attributes, as well as attitude towards the study would presumably remain constant.

The evaluation followed the same methodology used in the -ing evaluation described in Chapter 2: a featured-based evaluation where 4 professional native language speakers per target language judge grammaticality and accuracy through a binary question. Yet, we could not reuse the instructions without some modification, as we were no longer dealing with the translation of -ing words only. With the changes introduced by the new CL rules, the -ing words in implicit subjects were now subject and verb clauses and the -ing words in reduced relative clauses were *that* clauses. Only the rule aiming at introducing missing articles maintained the -ing words. Therefore,

we decided to delimit the structure to be judged by using highlighting. We highlighted the structures which resulted from the modifications applied (see Table 5.2) and asked the evaluators to judge the translation of the words highlighted in the source sentence (see Appendix H for instructions).

	Uncontrolled	Post-CL
CL-Rule 1	To create an XML file containing all parameters, use the /XML	To create an XML file that contains all parameters, use the /XML:
CL-Rule 2	The specified disk storage limit was reached when attempting to add a new revision to the desktop user data folder.	The specified disk storage limit was reached when you attempted to add a new revision to the desktop user data folder.
CL-Rule 3	Create cleaning job	Create a cleaning job

Table 5.2: Examples of rewrites for the new CL rules

We provided the evaluators with the source sentences and their MT output. In this round we did not include the PE version. The PE version we had was a rapid post-editing of the uncontrolled sentences. As a result, this version was very much influenced by the original source structure. We thought it would be confusing for evaluators to see acceptable structures in the new MT version modified by using alternative structures. For instance, the translation for "*When using...*" into Spanish was "*Al usar...*". However, after the CL rules were applied, "*When using...*" was changed to "*When you use...*". The translation for this structure was "*Cuando usted usa...*". The new translation is correct, yet, because the PE version belonged to the previous cycle, in the eyes of the evaluators it would look as if the post-editor would have changed it into "*Al usar...*". In order to avoid confusion, therefore, we decided not to include the PE version.

5.1.1.1 EVALUATOR AGREEMENT

Due to the smaller set of examples to be judged in this evaluation task, we decided to include 10% of sentences twice (20 examples) to be able to measure intra-rater agreement as well as the inter-rater agreement. The additional 20 examples were randomly chosen and the final 163 segments were randomly ordered in the evaluation file. We obtained inter-rater kappa scores of 0.754 for French, 0.429 for German, 0.609 for Japanese and 0.566 for Spanish, ranging from moderate for German and Spanish to substantial for French and Japanese. Of the 16 evaluators, the intra-rater agreement was 100% for 10 evaluators, 95% for three, 90% for two and 85% for one. All four

evaluators for Japanese reached 100% intra-rater agreement, and two out of the four evaluators for French, German and Spanish obtained this score. The results of these measurements confirm the reliability of the evaluation results by showing a good consistency in the judgements.

5.1.2 EVALUATION RESULTS

Let us first recall the results obtained in the -ing evaluation so as to then compare them to the results obtained in the present evaluation. By adding up the number of incorrect and inconclusive examples (these are the examples which would require some modification to upgrade to correct examples) we calculated the margins for improvement for the three groups per target language (see Table 5.3). The language with the highest number of incorrect examples addressed was French with 81 instances, followed by German with 57, Spanish with 40 and Japanese with 28. The structure and target languages which could benefit most were implicit subjects and reduced relative clauses for French and German. The maximum overall positive effect of the CL rules on the machine translation would mean an increase of 4.5% for correct -ing words for French, 3.17% for German, 1.56% for Japanese and 2.22% for Spanish.

Rule	Margin for improvement – Examples requiring improvement				
	French	German	Japanese	Spanish	TOTAL
Expand reduced relative clauses	29	27	17	23	96
Make implicit subject explicit	39	25	9	8	81
Insert article	13	5	2	9	29
TOTAL	81	57	28	40	

Table 5.3: Maximum margin for improvement for the CL rules

The human evaluation results showed a mixed effect on the machine translation output produced by the implementation of the new CL rules (see Tables 5.4-5.6). CL-Rule 1 improved the translation of 14 instances for French and 10 for Japanese, which obtained a statistically significant improvement/degradation ratio.⁵⁶ The German sample obtained one degradation, although 6 examples were now classified as *inconclusive* rather than *incorrect*. The worst performing language was Spanish, for which 13 degradations were reported. CL-Rule 2 was the most successful rule. It

⁵⁶ The statistical significance is reported based at a 95% confidence level.

improved the translation of 29 instances for French, 13 for German and 4 for Japanese. The improvement/degradation ratio was statistically significant for French and German, but not for Japanese. Spanish obtained 5 degradations. CL-Rule 3 was the worst performing rule, with degradations for German, Japanese and Spanish and only 5 improvements for French. The improvement/degradation ratio for French was not statistically significant. Overall, French saw the highest level of correct output with 48 instances; German improved the translation of 13 subordinate clauses but 3 of the articles inserted did not translate into better output. Japanese improved 14 incorrect translations but, similar to German, 2 of the articles inserted did not achieve a correct translation. Spanish was badly affected by the CL rules, obtaining 19 degradations.

	CL-Rule 1 - expand reduced relative clauses							
	FR before	after	DE before	after	JA before	after	ES before	after
correct	45	59	47	46	57	66	51	38
inconclusive	14	8	3	10	6	3	12	5
incorrect	15	7	24	18	11	5	11	30
total number of examples	74							

Table 5.4: Evaluation results for CL-Rule1

	CL-Rule 2 - make implicit subject explicit							
	FR before	after	DE before	after	JA before	after	ES before	after
correct	12	41	26	39	42	46	43	38
inconclusive	12	2	17	2	4	3	3	6
incorrect	27	8	8	10	5	2	5	7
total number of examples	51							

Table 5.5: Evaluation results for CL-Rule2

	CL-Rule 3 - insert article							
	FR before	after	DE before	after	JA before	after	ES before	after
correct	5	10	13	10	16	14	9	7
inconclusive	0	0	1	2	0	0	1	0
incorrect	13	8	4	6	2	4	8	11
total number of examples	18							

Table 5.6: Evaluation results for CL-Rule3

The aim of the CL rules was to eliminate the ambiguity introduced by the use of -ing words by making implicit information explicit. The ambiguity created by the -ing words was eliminated for most of the new source structures, in that no

gerund-participle was translated as a modifier, for instance. However, the RBMT was still not able to translate the -ing words correctly. This was due to other complexities present in the sentences. Remember that for the evaluation, only the new CL rules were applied to the evaluation sets, not a complete CL rule set. Therefore, the benefits of additional rules addressing other ambiguous structures could not be leveraged. Also, these results do not reflect the positive impact the rules might have for source text readers. The aim of CL is not only to improve machine translation but also to improve the readability and comprehensibility of source texts.

5.1.3 DEPLOYMENT INTO THE WORKFLOW

In order for the rules to be effective, it is of paramount importance that technical writers apply them efficiently. In this section we investigate the implications of deploying new rules. Particularly, we report on the experience of deploying them in a workflow which already works with a CL.

Let us very briefly consider the profile of the technical writer. Within an IT company, technical writers are the people responsible for creating the written content to be distributed with a product. According to White (1996), the skills expected from a technical writer, are the following:

- A bachelor's degree in communications, English (often with emphasis on writing in the professions), or in technical communication itself.
- Course work in computer science.
- Familiarity with a wide range of computer systems and capabilities.
- Familiarity with the mechanical and electronic systems about which one is writing.
- Knowledge of at least one programming language.
- Familiarity with desktop publishing and graphic design.
- Strong oral-communication skills.
- Good interviewing skills (in person or via telephone).
- Excellent command of written English; familiarity with report writing, letter and memorandum writing.
- Knowledge of and skills in information gathering.

(White 1996:14-15)

Overall, we could argue that trained technical writers are professionals with a strong competency in the English language and communication. Yet, training in technical writing is only recent and there is no guarantee that the writing teams always consist of well-trained professionals. As Hogg and Voss (2005) point out, before specific resources were ready to train technical writers, people from different fields entered the profession. Moreover, it must be taken into consideration that with the decentralisation of working groups, with more and more content writing performed in countries other than the US or the UK, trained professionals might be less available. Added to this is the fact that often writers are non-native speakers of English.

One of our first requirements was for technical writers to validate the performance of the new rules. Despite the high precision and recall values, it was important that the writers were aware of the proposed changes to the rule set and that they felt comfortable with their implementation. With the current state of the profession in mind, and aware that Symantec teams are spread around the globe, we approached the editing team. The team is based in the US, and is responsible for different writing teams across the world. The editors, constantly working with the technical writers, are aware of their training and knowledge, and know their needs.

The editors agreed to collaborate on this task. We sent them a set of 100 randomly chosen flagged sentences per rule for evaluation together with the explanation of the focus of the new rules. Additionally, they were also given the option of deploying the rules in a test environment and evaluating them on their own data. By providing these two options we aimed at offering different evaluation approaches. Whereas the former was a more controlled option, the latter reflected a real-life scenario.

The editors raised concern over the handling of terminology, asking for some modification, and gave their approval for the deployment of the rules. Their only resistance arose from the fact that the rules did not take approved terminology into account. For example, if the term *paging file*, included in the term bank, appeared in an index without a determiner, acrolinx IQ™ would flag it because it violated the rule stipulating that articles should precede singular noun phrases. We also noticed some confusion, probably created by the change of approach taken with the new rule. Previously, writers were used to eliminating flagged -ing words. Now, we were asking them to check specific grammar around the -ing words, not to eliminate them. This

point was clarified and the deployment proceeded. Yet, the rules were tuned to discard all flagged instances containing approved terminology. We were aware that this decision would decrease recall but it was considered a compromise worth making for a successful implementation.

Secondly, we required feedback on the integration of the new rules within the existing CL rule set. Two options were possible. On the one hand, we could include the new rules within the already existing rules dealing with articles, implicit subjects and relative pronouns. On the other hand, we could group them in an -ing-specific rule. We left this decision to the editors. They agreed that the second option was more appropriate. Editors argued that dealing with two grammatical issues at the same time, the issue with the -ing words and the implicit subjects, for instance, would be too demanding on the writers (Idoura).⁵⁷ Once again we noticed that the habit of seeing -ing words as a problem to be removed was creating confusion in the new implementation. Our research showed that certain -ing subcategories and combinations were problematic, not -ing words in general. However, from the arguments used by the editors it was clear that they still viewed -ing words in general as problematic items they had to eliminate. The misunderstanding was again cleared up. However, the rule was implemented according to the second option whereby a rule specifically addressing -ing words in combination with specific grammatical structures was created. As a first step to break the habit of viewing -ing words as being problematic overall, it was decided that the name of the rule would be changed from *avoid_ing_words* to *disambiguate_ing_words*. Note also that for the sake of consistency, we modified the rules dealing with the insertion of articles, implicit subjects and relative pronouns so that they would not flag any instances of -ing words.

We also needed to consult with the end users on the content of the help files accompanying the new rules. Editors were presented with a quick access to the help file. The files contained a brief description of the error and suggested changes, along with examples of incorrect sentences and their corrected counterparts. The editors requested the language in the help files to be simplified arguing that "*Writers are not grammarians and may have limited familiarity with English grammar (which is why*

⁵⁷ E-mail communication from A. Idoura, Department of Structured Content Strategies, received on 15 August 2008, compiling responses from 3 editors.

they are using acrocheck in the first place)” (Idoura).⁵⁸ They also added that writers mostly rely on the correct and incorrect examples when re-writing and suggested adding more examples. Following the editors’ advice, we re-wrote the explanations of the violations. We re-used existing phrasing which writers were familiar with as much as possible and only added new simple grammatical concepts when necessary. We completed the modifications by adding extra examples to cover the most frequent errors.

5.2 AUTOMATIC SOURCE RE-WRITING

We wrote 3 rules in order to test this technique. Rule 1 transforms the -ing words at the beginning of titles into imperatives. This rule was tested for Spanish because the RBMT system can be configured to translate imperatives into Spanish infinitives. Titles starting with infinitives were classified as correct by evaluators in the -ing evaluation. Rule 2 transformed the -ing words at the beginning of titles into nouns followed by the preposition *of*. This rule was expected to have a positive effect on the four target languages and therefore applied to French, German, Japanese and Spanish. The rules were applied to the instances grouped under the title category in the evaluation corpus and evaluated. Rule 3 introduced the subject *you* into implicit *when* + *ing* structures. This rule was tested for French and Japanese.

5.2.1 EVALUATION SET-UP

We were no longer in a position to reproduce the same evaluation setting as the first -ing evaluation and the CL rule evaluation. The RBMT system had since been upgraded and only three out of the four original evaluators per target language were available. Therefore, we decided to use a different evaluation type: ranking of different translations of -ing words. As described in Chapter 2, this evaluation type has both advantages and disadvantages. The advantages are that the evaluation is reported to be less demanding in terms of cognitive effort because the overall quality of the evaluated unit must not be decided. As a result, the evaluation is performed faster. On the other hand, by asking which translation is better, we would not know whether the output was actually grammatical and accurate or not. Yet, this evaluation aims at testing whether

⁵⁸ E-mail communication from A. Idoura, Department of Structured Content Strategies, received on 15 August 2008, compiling responses from 3 editors.

improvement in translation quality is achieved by using a particular pre-processing technique and therefore it was considered sufficient.

We machine translated the original and automatically modified sentences obtained after applying the rules to the sentences classified under titles and adverbials of time introduced by *when*. Two outputs were obtained. MT1 belongs to the original source. MT2 belongs to the automatically modified source. 4 professional native language speakers per target language (three from the previous human evaluations and one new for each target language) were asked to judge which output contained the best translation of the -ing word in terms of grammaticality and accuracy. In order to focus evaluators on the unit of evaluation (-ing words) we again highlighted the relevant -ing word in the original source sentences. The RBMT system transferred the highlight to the word(s) in the target sentence. We could not automatically obtain the highlight for MT2, as the new source did not have any highlight and therefore MT2 was presented without any highlighting. Evaluators were presented with the original source sentence, MT1 output and MT2 output. Additionally, and in order to reduce cognitive effort, we presented the subcategories in groups. Evaluators were provided with three possible answers: MT1, SAME and MT2. When the first and third options were selected, it would be clear which of the outputs was better. The selection of SAME would have different indistinguishable meanings: the translation of the -ing word is the same, both translations are equally bad and both translations are equally good.⁵⁹ The aim being to report whether improvement was achieved, we considered that asking for this level of detail from evaluators was not necessary (See Appendix I for guidelines).

5.2.2 EVALUATION RESULTS

In order to calculate the effect on the translation, we added up the score obtained for each -ing word. Only the examples where at least three evaluators agreed were used to report the results. By doing that, we were able to use 83% and 87% of the examples for Spanish for Rule 1 and Rule 2 respectively; 81% and 80% for French for Rule 2 and Rule 3 respectively; 70% for German for Rule 2; and 44% and 76% for Japanese for Rule 2 and Rule 3 respectively (see Useful Examples vs. Total Examples in Tables

⁵⁹ We checked for MT1 and MT2 sentences that were the same. These sentences were not included in the evaluation. They were later recovered to report overall results.

5.5-5.11). The inter-rater kappa scores obtained for each rule and target language during this evaluation varied from good to very good for Spanish, fair and no agreement for French, good for German and fair and no agreement for Japanese (see Tables 5.7-5.13 for scores).

5.2.2.1 SPANISH

ASR-Rule 1 transformed -ing words at the beginning of titles into imperatives. The output obtained from this new source text was judged as better for 32% of the examples. The output was judged the same almost half of the time. MT1 was considered better for 20% of the examples. The result for ASR-Rule 2 was more positive for MT2, which was judged better 70% of the time. MT1 was judged better only 2% of the time and the same quality reported for 27% of the examples. We observed that both rules improved the translation quality of sentences. The improvement/degradation ratio was statistically significant for both rules.

ASR-Rule 1	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	97	227	157	481	557	0.761
%	20.17	47.20	32.64	83.36		

Table 5.7: Results for rule transforming -ing titles into imperatives for Spanish

ASR-Rule 2	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	11	138	357	506	557	0.825
%	2.17	27.27	70.55	87.69		

Table 5.8: Results for rule transforming -ing titles into nouns for Spanish

5.2.2.2 FRENCH

ASR-Rule 2 transformed -ing words at the beginning of titles into nouns. For this rule, MT1 was judged better for 7% of the examples. The output was reported to be of the same quality 22% of the time. MT2 was judged better 70% of the time. Similarly to Spanish, the improvement/degradation ratio obtained with nominal titles was statistically significant, with the degradation, although 4 points higher, still very low.

ASR-Rule 2	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	34	102	320	456	557	0.448
%	7.46	22.37	70.17	81.87		

Table 5.9: Results for rule transforming -ing titles into nouns for French

ASR-Rule 3 transformed *when + ing* structures into explicit clauses with the subject *you*. This rule greatly favoured the translation quality of the -ing words

(improvement/degradation ratio statistically significant). No degradations were reported and MT2 was judged better for 96% of the examples.

ASR-Rule 3	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	0	1	28	29	36	-0.015
%	0	3.45	96.55	80.56		

Table 5.10: Results for rule transforming -ing adverbials introduced by *when* for French

5.2.2.3 GERMAN

ASR-Rule 2 transformed -ing words at the beginning of titles into nouns. For this rule, MT1 was judged better for 17% of the examples. The output was reported to be of the same quality 22% of the time. MT2 obtained 60% of the scores (improvement/degradation ratio statistically significant improvement). The degree of degradation is higher than that for the other three target languages.

ASR-Rule 2	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	69	88	234	391	557	0.657
%	17.65	22.51	59.85	70.20		

Table 5.11: Results for rule transforming -ing titles into nouns for German

5.2.2.4 JAPANESE

Japanese scores were very similar to Spanish and French for ASR-Rule 1. Degradation was minimal at 5% and improvement was high at 75% (improvement/degradation ratio statistically significant).

ASR-Rule 2	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	23	53	323	427	557	0.427
%	5.39	18.97	75.64	76.66		

Table 5.12: Results for rule transforming -ing titles into nouns for Japanese

ASR-Rule 3 transformed *when* + *ing* structures into subordinate *you* clauses. Whereas degradations were not reported, the majority of the cases (81%) were judged as having the same quality. 18% of the instances were better for MT2 (statistically significant improvement/degradation ratio). The low number of examples for which at least three evaluators agreed showed that the evaluators found it hard to judge the quality of the MT output. The improvement

ASR-Rule 3	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	0	13	3	16	36	0.092
%	0	81.25	18.75	44.44		

Table 5.13: Results for rule transforming -ing adverbials introduced by *when* for Japanese

5.2.3 DEPLOYMENT INTO THE WORKFLOW

The applications of automatic source re-writing were performed by using Regex and the controlled language checker. Should a localisation workflow have a CL checker in place, adding an automatic check before the source text is to be machine translated would be possible. A full-fledged CL checker would already support the document formats in which the source text is created. A possible problem would be the processing time required to analyse and transform the source text.

Applying rules based on Regex, however, would require some considerations. Regex can be applied directly into the text using a text editor that supports them or integrated within a script. Therefore, it is of paramount importance that the source text format is compatible with Regex. If the source text is written in XML format, for instance, the modifications could be applied directly. Although in this case the fact that tags could occur within the strings to be modified would have to be taken into consideration. However, if the source text is written in a proprietary text format such as Microsoft Word's an extra step to transfer the text into a compatible format without losing formatting information, would have to be considered.

5.3 GLOBAL SEARCH & REPLACE

Global S&R was tested using two different rules. GSR-Rule 1 removed articles preceding infinitive verbs at the beginning of titles. This rule was written to address Spanish MT output. GSR-Rule 2 and GSR-Rule 3 transformed the word classes of the objects of catenative verbs. Both Spanish and French examples were judged as incorrect for this type of problems. Because Global S&R is a post-processing technique which is applied to the MT output, language-dependent rules were written. GSR-Rule 2 was written to specifically target Spanish and GSR-Rule 3 for French.

5.3.1 EVALUATION SET-UP

This evaluation followed the same methodology used for automatic source re-writing, the only difference being that in order to test a post-processing technique, evaluators were presented with the source sentence, the original RBMT output (MT1) and the output obtained after applying the Global S&R rules to the original RBMT output (MT2). Four professional target language speakers (three out of four for each target language had participated in the first -ing evaluation) were asked to judge which version, MT1 or MT2, included a better translation of the -ing word highlighted in the source in terms of grammaticality and accuracy. Should any case occur where the quality remained the same, the option “SAME” was provided (see Appendix I for guidelines).

5.3.2 EVALUATION RESULTS

Similarly to automatic source re-writing, we report the results for the examples where at least three out of the four evaluators agreed on the quality. By doing so, we report the results for 72% of the examples for GSR-Rule 1, 20% for GSR-Rule 2 and 83% for GSR-Rule 3. We will come back to the high level of inconclusive examples for GSR-Rule 2 in the section below. Kappa scores to measure inter-rater agreement were calculated for each of the rules. GSR-Rule 1 obtained 0.473, a fair agreement. GSR-Rule 2 obtained a negative score. For GSR-Rule 3, a score of 0.269 was obtained; a slight agreement.

5.3.2.1 SPANISH

GSR-Rule 1 transformed the word class of catenative verbs including the prepositions, should they be incorrect, for Spanish. The original RBMT output for the translation of the -ing word was considered of better quality (in terms of grammaticality and accuracy) 7.7% of the time. The quality was judged to be the same 30% of the time. The output which included the effect of GSR-Rule 1 was evaluated as better for 62% of the examples, proving a statistically significant improvement/degradation ratio.

GSR-Rule 1	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	1	4	8	13	18	0.473
%	7.7	30.8	62	72.22		

Table 5.14: Results for the rule removing articles in front of infinitive verbs at the beginning of titles for Spanish

GSR-Rule 2 removed articles preceding infinitive verbs at the beginning of titles for Spanish. As mentioned, out of the 15 examples tested for this rule, at least three evaluators only agreed in 3 cases. A closer look at the examples showed that two evaluators constantly judged the translations where the articles were removed as better, whereas two evaluators constantly judged them as being of the same quality. Only in 3 occasions they all agreed that the quality was the same. This was due to additional ungrammaticalities being present in the translation, such as the use of reflexive pronouns.

GSR-Rule 2	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	0	3	0	3	15	-0.333
%	0	100	0	20		

Table 5.15: Results for the rule transforming the word class of the objects of catenative verbs for Spanish

5.3.2.2 FRENCH

GSR-Rule 3 transformed the word class of catenative verbs including the prepositions, should they be incorrect, for French. Of the 15 examples for which at least three evaluators agreed on the result, 80% were judged to be of better quality after applying GSR-Rule 3 and 20% remained at the same level (statistically significant improvement/degradation ratio).

GSR-Rule 3	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	0	3	12	15	18	0.269
%	0	20	80	83.33		

Table 5.16: Results for the rule transforming the word class of the objects of catenative verbs for French

5.3.3 DEPLOYMENT INTO THE WORKFLOW

Global S&R was applied using Regex, similarly to ASR-Rule 1 and ASR-Rule 2 in automatic source re-writing. The considerations regarding the file formats in order to implement Regex were discussed in section 5.2.3. What differentiates the deployment of Global S&R is the workflow stage in which it must be applied: after the MT system generates the target output and before this output is transferred to the following stage. The specific deployment, therefore, will depend on the possibilities and requirements of each workflow, the file formats used and the possibilities of inserting scripts.

5.4 STATISTICAL POST-EDITING

In Chapter 4, an SPE engine was built using the Moses decoder. In order to test the effect on the final translation quality, a statistically representative set of sentences containing -ing words was post-edited. This section presents the human evaluation set-up and results, as well as some considerations for the deployment in a localisation workflow.

5.4.1 EVALUATION SET-UP

The human evaluation type performed for SPE was similar to the one used for automatic source re-writing and global search & replace: ranking of two different translations MT1 and MT2 (see Appendix I). MT1 contained the RBMT translations. MT2 contained the RBMT+SPE translation output. However, due to the general approach encompassed by SPE whereby modifications could be applied for the whole sentence and were not targeted to the -ing words, two questions were asked. Evaluators were asked to select which MT output contained the best -ing word translation (or whether the quality was the same) and which MT output was, at a sentence level, a better translation of the source sentence. This information was deemed necessary to obtain a full picture of the effect the SPE module had on the evaluation set. Restricting the evaluation to the -ing words would have been misleading. The -ing word might have improved in a sentence, but the SPE module could have decreased the level of translation quality elsewhere.

The evaluators were presented with the source sentences where the -ing words to be evaluated were highlighted, the RBMT output of each sentence with the translation of the -ing words highlighted and the RBMT+SPE output with no highlighting. Additionally, the sentences, which belonged to different subcategories of the -ing classification, were presented randomly.

5.4.2 EVALUATION RESULTS

The evaluation results presented in this section were obtained from the examples in which at least 3 evaluators agreed on the quality. We were able to report results for 76-98% of examples for the sentence and -ing word evaluations separately and for 75% for cross-tabulations. The inter-rater agreement kappa scores obtained for this

evaluation were fair and moderate, with a poor sentence-level agreement for French (see specific scores in Tables 5.17-5.24).

5.4.2.1 SPANISH

The human evaluation shows that at a sentence level MT2 scored best 25% of the time and the degradation was 10%, showing statistically significant improvement/degradation ratio. The highest percentage, 64%, belongs to the evaluation category which reports equal sentence quality.

Sentence	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	25	161	63	249	317	0.608
%	10.04	64.66	25.30	78.55		

Table 5.17: Results for sentence-level quality for Spanish

The results for the translation of the -ing words report 8% improvement and 3% degradation using the SPE module. The difference results in statistically significant improvement/degradation ratio. 88% of the -ing words, however, display equal quality.

ING	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	9	260	26	295	317	0.656
%	3.05	88.13	8.81	93.06		

Table 18: Results for -ing words-level quality for Spanish

A cross-tabulation of the -ing constituent and sentence-level results shows that for the cases where MT1 obtains a better rank, the -ing constituents are reported to be of the same quality 83.33% of the time. No cases are reported where MT1 is better at sentence-level and MT2 is better at -ing word level. Similarly, when MT2 is better for sentence-level quality, only 1.7% of the examples were judged to have a better quality for MT1, with 41% judged to be better in MT2. Yet, the majority of the examples fall into the “same” quality category.

ING Sentence	MT1 Best	SAME	MT2 Best	Useful Examples	%
MT1 Best	2	22	0	24	10.21
SAME	1	146	6	153	65.10
MT2 Best	1	39	18	58	24.68
Useful Examples	4	207	24	235	74.13
%	1.70	88.08	10.21	74.13	

Table 5.19: Cross-tabulation of -ing word and sentence-level results

5.4.2.2 FRENCH

The human evaluation for French shows an even higher percentage of both sentences and -ing words judged as having the same quality, 88% and 95% respectively. The improvement degradation ratio for the sentence-level evaluation is 6% against 5%. At the -ing word-level the same percentage of improvement and degradation was obtained. SPE did not bring statistically significant improvement/degradation ratio for this TL.

Sentence	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	14	214	15	243	317	0.255
%	5.76	88.06	6.17	76.66		

Table 5.20: Results for sentence-level quality for French

ING	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	8	295	8	311	317	0.552
%	2.57	94.85	2.57	98.11		

Table 5.21: Results for -ing word-level quality for French

The cross-tabulation shows that sentence-level and -ing word-level quality correlated when either MT1 or MT2 were selected as best output. 1% to 2% of -ing words displayed a better MT1 or MT2 translation when the sentence-level translation was judged of similar quality. 4% to 5% of sentences were classified as having better MT1 or MT2 quality when the -ing words were reported to be of equal quality.

ING Sentence	MT1 Best	SAME	MT2 Best	Useful Examples	%
MT1 Best	3	11	0	14	5.81
SAME	2	207	4	213	88.38
MT2 Best	0	13	1	14	5.81
Useful Examples	5	231	5	241	76.02
%	2.07	95.85	2.07	76.02	

Table 5.22: Cross-tabulation of -ing word and sentence-level results

5.4.2.3 GERMAN

The human evaluation for German shows a lower percentage of examples being judged as having the same translation quality at sentence level. However, we observe that the percentage of degradations is higher than the improvements introduced by SPE, 25% against 23%.

Sentence	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
----------	----------	------	----------	-----------------	----------------	-------

Examples	62	128	58	248	317	0.668
%	25	51.61	23.39	78.23		

Table 5.23: Results for sentence-level quality for German

At a -ing word-level, the improvement degradation ratio is 6% to 5% (improvement/degradation ratio not statistically significant). The percentage of examples judged to have the same translation quality is similar to the percentages obtained for French and Spanish, at 88%.

ING	MT1 Best	SAME	MT2 Best	Useful Examples	Total Examples	KAPPA
Examples	15	263	18	296	317	0.587
%	5.07	88.85	6.08	93.37		

Table 5.24: Results for -ing word-level quality for German

A cross-tabulation of the results shows that when the translation of the -ing word is better for MT1, the sentence-level quality is also better for MT1. Similarly, when the the sentence-level quality is the same for MT1 and MT2, the translation of the -ing word is of equal quality for MT1 and MT2. When MT1 and MT2 are considered better at a sentence-level, 75% and 70% of the examples are judged to be of the same quality for the translation of the -ing word.

ING Sentence	MT1 Best	SAME	MT2 Best	Useful Examples	%
MT1 Best	12	43	2	57	24.25
SAME	0	125	0	125	53.19
MT2 Best	0	37	16	53	22.55
Useful Examples	12	205	18	235	317
%	5.10	87.23	7.66	74.13	

Table 5.25: Cross-tabulation of -ing word and sentence-level results

5.4.3 DEPLOYMENT INTO THE WORKFLOW

A number of technical considerations are required in order to deploy an SPE module in a localisation workflow. Once the benefits to the translation quality are measured, the cost, both in terms of processing capacity, time and formatting, must be quantified. By using the equipment described in 4.2.2.1, the translation of 385 sentences took around 20 minutes. A balance between an acceptable time and processing capacity would have to be found. Also, at the moment formatting is not supported by the engine. Firstly, the same case described in the automated source re-writing arises with the issues of

proprietary document formats. Secondly, open formats would require some workaround. An XML file, for instance, could be input to the engine. However, the system must be trained to handle the XML tags. The option exists of removing tags before the text is post-edited. Later, the tags would have to be re-inserted. A second option would be to train the SPE engine with XML-tagged documents. However, the additional variation introduced would result in the need for larger corpora for training.

5.5 CHAPTER SUMMARY

The impact on the translation quality of pre- and post-processing techniques was investigated in this Chapter. We performed four human evaluations to observe the variation in translation introduced by each of the four techniques described in Chapter 4 (see Table 5.26 for a summary of improvement results). For controlled language rules, a correct machine translation of the -ing words was not always achieved. The automated source re-writing for titles transformed to begin with nouns improved the translation output significantly, with 60% improvement for German and 70% for French and Spanish and 80% for Japanese. What is also important is that the degradation introduced by the rule was minimal at 2-7% for French, Spanish and Japanese but was higher for German at 17%. The rule dealing with implicit subordinate clauses achieved a 96% improvement and no degradations for French. Japanese did not benefit from this rule to the same degree, obtaining only 18% improvement. The Global S&R displayed an improved translation of 62% and 80% when dealing with catenative verbs for Spanish and French, respectively. The rule dealing with articles preceding infinitives did not prove to improve translation quality. Yet, no degradations were reported. In general, we observed that rule-based techniques offered the possibility of targeting efforts for particular -ing word subcategories and issues. Additionally, the results show that when writing a MT-specific rule with automated modifications, the degradations introduced were minimal.

Contrary to the results reported in the field, our SPE engine did not achieve considerable improvement at either sentence or -ing word-level for any of the target languages. After building the SPE engine following the size/quality of the training corpus common in the field, our results show minimal improvement for Spanish, at

both sentence and -ing word-level, no improvement for French and degradation at sentence-level and no improvement at -ing word-level for German.

Technique		ES	FR	DE	JA
CL	CL-R1	-52%	48%	-4%	53%
	CL-R2	-62%	74%	52%	44%
	CL-R3	-22%	38%	-60%	-100%
ASR	ASR-R1	32%	n/a	n/a	n/a
	ASR-R2	70%	70%	60%	80%
	ASR-R3	n/a	96%	18%	n/a
GS&R	GS&R-R1	62%	80%	n/a	n/a
	GS&R-R2	0%	n/a	n/a	n/a
SPE		8%	0%	6%	n/a

Table 5.26: Overall improvement results for the tested -ing words

CHAPTER 6

CHAPTER 6: CONCLUSIONS

6.1 OBJECTIVES

This dissertation aimed at answering two main questions. The first sought the identification of the -ing words which are problematic for RBMT systems when translating procedural and descriptive IT texts into French, German, Japanese and Spanish. This was answered by classifying -ing words extracted from a representative corpus of IT descriptive and procedural texts using a functional scheme (Izquierdo, 2006) and evaluating their machine translation.

With the human scoring obtained, we sought to test the correlations between automatic metrics and the human evaluation. We noted that correlations were good, with some metrics appearing to be more sensitive to certain TLs over others.

The second question sought the exploration of techniques to improve the translation of the -ing words. This was accomplished by analysing the characteristics of four techniques in terms of the language (SL or TL) they work on, the source of the error (shared or TL-specific), the language resources they require for a successful implementation, the need for human interaction, the purpose of the end-product and the scope of the modifications; and by choosing problematic -ing word subcategories to test the techniques' feasibility. Two measurements were necessary to report on the performance of the techniques. Precision and recall were calculated to measure the performance of the rules and a human evaluation was performed to assess machine translation improvement.

6.2 FINDINGS

While performing the classification of -ing words in our corpus, we were able to get an insight into the use of -ing words within IT procedural and descriptive texts. The concentration of -ing words in our corpus was 2%, coinciding with concentrations reported by Biber (1988) and found in the BNC for similar documents. This suggests that the degree of use of -ing words in this specialised corpus is not significantly different from general language corpora. In addition, we saw that the most populated category was that of titles, with 25% of the share. Characterisers followed, with 23%, mainly used as pre-modifiers. The category of Adverbials, with a 19% occurrence rate,

contained most -ing words as heads of mode (*by + ing*), purpose (*for + ing*) and temporal (*when + ing*) phrases. The use of Referentials and Progressives was restricted to 7% and 6% respectively. By selecting, implementing and customising (to a small degree) a proposed classification for -ing words, we have validated this classification and shown its applicability to a text genre different from the one for which it was initially developed.

One of the main findings of this dissertation has to do with the machine translation quality of -ing words. The results from the human evaluation showed that the majority of -ing words were correctly handled by the RBMT system – 72% for German, Japanese and Spanish, 52% for French. This shows that, despite the theoretical difficulties involved in the translation of this feature, the current development stage of the RBMT system used enables the successful resolution of a considerable proportion of -ing words. However, at 72%, that still leaves 28% of occurrences rated as problematic, which could represent a significant correction effort if one is translating millions of source words per annum. Overall, Titles and Characterisers were among the best performers for German; Titles were the worst performers for French. Progressives varied across languages obtaining the worst overall results for Japanese and the best for French. Referentials, in turn, were among the worst performers for all four languages. This demonstrates that the problems generated while machine translating -ing words are somewhat TL-specific.

Also, despite the fact that considerable effort had been put into coding UD's prior to the research project, we found that incorrect terminology played a significant role in translation of -ing words marked as "incorrect". This outcome highlights the importance of terminology management for the production of machine translation quality. Terminological accuracy in RBMT systems is controlled through the customisation of UD's with previously selected terms. Given the influence of terminology, a revision of terminology extraction processes is proposed. Additionally, the increasing encoding possibilities offered by UD's, which directly interact with the RBMT system, in terms of contextualisation and linguistic information, provide a promising area for research, where the effect on disambiguation could be measured.

The comparison between human and automatic metrics' evaluation showed moderate agreement (0.42-0.75) at a feature-level and almost perfect agreement

(0.86-0.98) when calculating a human-score aggregated correlation. This is in the same range of agreement that is reported for evaluations performed at sentence level within the field. Therefore, the results seem to support the suggestion that, despite the fact that automatic metrics are optimised for sentence or textual units, shorter segments obtain similar correlations with human scorings.

The analysis of the techniques to improve the translation of -ing words revealed their individual requisites as well as the types of issues which would benefit most from their implementation. For the pre-processing stage, controlled language was found to be effective in improving the original text by eliminating ambiguities and ungrammatical structures, which in turn facilitates a more accurate analysis by the RBMT system. The CL rules that were tested revealed mixed results with regard to translation improvement. French performed well with the three new rules, improving 60% of the incorrect instances. German and Japanese also improved in translation quality by 23% and 50%, although a low degradation level was also observed (5% and 7% respectively). Spanish was the language with the poorest response, showing 33% degradations. We also found that specific CL rules applied in isolation of other rules in the set can lead to output that is not of optimal quality. CL rules need to be applied as a set in order to gain the most advantage. This resulted in the re-written -ing words not always being translated correctly due to the combination of additional complexities in the source sentences.

Automatic source re-writing proved an effective RBMT-specific and language-pair-specific technique for increasing machine translatability. Its main strength lies in not having to adjust to grammatical constraints when applying modifications. The source text can be automatically modified to better suit the MT system, usually converting the source text into a form closer to the expected target text. Both CL and automatic source re-writing reduce source text complexity for RBMT systems, which should result in improved translation quality. However, no matter how high the degree of determinism is in RBMT systems, when working with a proprietary system - in a black-box environment -, predicting translation performance is still challenging. In the case of automatic source re-writing, the translation improvement was significant, at 60% to 90%. One of the main findings was that the degradation

resulting from the modifications was minimal. Given the significant improvement seen by us, we expect this method will receive further attention in the future.

When issues cannot be tackled by modifying specific source structures and are target-language specific, post-processing techniques can be used to modify the translated text. Global search & replace was observed to have the potential to solve TL errors by using manually crafted rules. Although feasible, we showed that limitations for certain types of modifications existed by using Regex only. In fact, both automatic source rewriting and global search & replace would benefit from the inclusion of linguistic information, deeper structural analysis and a linguistically-informed generation module, or at least more flexible replacement capabilities. We would highlight this as a potential area for research and development into the future.

Statistical post-editing was the data-driven technique studied where transformations were automatically learnt and where general instead of -ing word subcategory-specific modifications were applied. The training of the engine required large volumes of aligned MT and reference sentences. The results obtained showed little improvement, and even degradation depending on the target language. This is not in line with the results reported in the field where, using same size and domain-specific training data, improvements over the baseline system are claimed.

6.3 REVIEW OF METHODOLOGIES

One of the strengths of this research is that it has evaluated the machine translation of a particular linguistic feature, i.e. -ing words, in real-life IT documents. In order to identify which particular subcategories were problematic for the RBMT system, we opted for a feature-based evaluation. Identifying an -ing word and providing it in context was straightforward. However, identifying the exact translation segment that was relevant for evaluation was not. Defining which information was condensed within the -ing words apart from the lexical meaning was challenging. As a result, establishing which word(s) contained such information in the target text for evaluation was difficult. The exact translation of the -ing words for each of the subcategories was decided through the mapping used by the MT system as well as additional human decisions to include or exclude the effects of linguistic features in the near context. This enabled us to obtain consistent judgements from a well-defined evaluation.

Additionally, the validity and reliability of the results were further ensured by the selection of evaluators and tests to measure the agreement rates between them. We believe that the use of the challenging feature-based evaluation and the strict controls implemented during evaluation contribute to the strengths of this study. Real-life instances were selected, evaluated and analysed, and the problematic -ing word subcategories identified.

The second part of this research focused on techniques to improve the problematic -ing word subcategories. In order to test their feasibility and efficiency, we selected a number of -ing subcategories which, given their characteristics, were best suited for each technique. For the rule-based targeted techniques, i.e. controlled language, automatic source re-writing and global search and replace, we reported on the general strengths and limitations, as well as precision and recall measurements obtained for the rules created. Secondly, we evaluated the new machine translation output using professional evaluators and reported the effect on translation quality. This allowed us to draw conclusions on the potential of the techniques to improve the machine translation quality of a number of -ing word subcategories. We reported the challenge involved in the initial implementation of the techniques but we did not measure the exact implementation effort required. This would be necessary when performing an overall cost-effectiveness analysis of each technique on top of the reported performance measurement. We have established the baseline of conditions and advantages of each technique and proved their potential.

While we have focussed on one specific linguistic feature throughout, we hope that the details of our methodology could be applied equally to any other linguistic feature one might be interested in. In that respect, we have contributed to the general methodology for performing feature-based evaluations of both MT output and solutions.

As mentioned in our Introduction, the special setting for this research (i.e. academic/industry collaboration) meant that the research was to some extent driven by the needs of the industrial partner. Rather than seeing this as a problem, we see the collaboration as a strong, positive aspect of the research as it enabled the use of “real” data, professional linguists and a commercial MT system to investigate an issue that

had been previously seen as a “problem” and to implement and test not one, but several, potential solutions to that problem.

6.4 FUTURE CHALLENGES AND RESEARCH OPPORTUNITIES

Due to its broad coverage, this thesis identified several directions for future research. First of all, the findings of the present study are based on one RBMT system, i.e. SYSTRAN. Further tests to examine the performance of RBMT systems regarding -ing words would be necessary to confirm that the reported issues are shared among RBMT systems and not the result of the particular development platform of SYSTRAN.

Also, the evaluators hired to judge the machine translation of the -ing words were professional translators. They were suitable because we set grammaticality and accuracy of texts as evaluation attributes. Given that the texts aim at instructing IT users about how software products function, it would be interesting to test the usefulness and usability of machine-translated text in general through studies of end users’ ability to comprehend and use the MT output to perform tasks.

The weaknesses of performing a human evaluation range from the large budget, necessary if we are to avail of professionals, to the exhaustive evaluation design required. The cognitive effort of evaluators needs to be considered and balanced with the minimum scope required to ensure an informative and trustworthy response. The task has to be strictly designed (and guidelines provided) to prevent misinterpretations and to be able to compare results across evaluators. We explored the use of a number of string-based automatic metrics as alternative to human evaluation, focusing on their usefulness at a subsentential level. The need for further research in this area was clear from the results. First, additional research would be required to confirm the success of automatic metrics when evaluating grammatical features, and whether certain languages correlate better with specific metrics. Also, it would be interesting to examine the impact the translation quality has on the reported correlations. Research effort could also be channelled into investigating what correlations between automatic metrics and human judgements suffice for an effective deployment of the automatic metrics.

Finally, we would like to mention the integration of source parsing, tagging and morphological analysis, and generation capabilities that could be developed for the techniques for improvement described in this research. We examined and tested a number of techniques for improvement “out-of-the-box”. During the analysis and particularly during the implementation procedures, we identified limitations related to the availability of linguistic information. We argued that additional capabilities, which are already being developed within the NLP community, would increase their efficiency. The path for further research in these areas, based on the -ing word example, is therefore set.

6.5 CLOSING REMARKS

This research has dispelled the myth that -ing words are particularly problematic for MT for four target languages. It has identified the problematic -ing word subcategories. It has built on current evaluation methods to establish a feature-based approach and performed a large-scale multilingual human evaluation, complemented by automatic metrics. It has contextualised and assessed the feasibility and effect of several pre- and post-processing solutions, initiating the integration of innovative technology into the automated translation industry.

REFERENCES

- Adriaens, G. & Macken, L. 1995. Technological evaluation of a controlled language application: precision, recall, and convergence tests for SECC. IN: *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 1995)*, Leuven. pp.123-141.
- Adriaens, G. & Schreurs, D. 1992. From COGRAM to ALCOGRAM: Toward a controlled English grammar checker. IN: *Proceedings of the 14th Conference on Computational Linguistics (COLING 1992)*. Nantes, France. pp.595-601.
- Aiken, M.W. & Wong, Z. 2006. Spanish-to-English translation using the web. IN: *Proceedings of the 37th Annual Conference of the South West Decision Science Institute*, Oklahoma. pp.161-166.
- Akiba, Y., Imamunra, K. & Sumita, E. 2001. Using multiple edit distances to automatically rank machine translation output. IN: *Proceedings of the MT Summit VIII: Machine Translation in the Information Age (MTS VIII)*, Santiago de Compostela, Spain.
- Albrecht, J. & Hwa, R. 2007. A re-examination of machine learning approaches for sentence-level MT evaluation. IN: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic.
- Allen, J. 1999. Adapting the concept of “translation memory” to “authoring memory” for a controlled language writing environment. IN: *Proceedings of the 21th Conference of Translating and the Computer (ASLIB 1999)*, London.
- Allen, J. 2003. Post-editing. IN: H. Somers (ed.) *Computers and Translation: A Translator’s Guide*. Amsterdam: John Benjamins Publishing Company. pp.297-1317.
- Allen, J. & Hogan, C. 2000. Toward the development of a postediting module for war machine translation output: a controlled language perspective. IN: *Proceedings of the 3th Workshop on Controlled Language Applications (CLAW 2000)*, Seattle, Washington. pp.62-71.
- Almqvist, I. & Sågval-Hein, A. 1996. Defining ScaniaSwedish – a controlled language for truck maintenance. IN: *Proceedings of the 1st International Workshop on Controlled Language Applications (CLAW 1996)*, Katholieke Universiteit Leuven, Belgium.
- Almqvist, I. & Sågval-Hein, A. 2000. A language checker of controlled language and its integration in a documentation and translation workflow. IN: *Proceedings of the 22nd Conference of Translating and the Computer (ASLIB 2000)*, London.
- Aranberri, N. & Roturier, J. 2009. Comparison of alternatives to strict source control: a case study with -ing words. IN: *Pre-Proceedings of the Workshop on Controlled Natural Language (CNL 2009)*. Marettimo Island, Italy.
- Areta, N., Gurrutxaga, A., Leturia, I., Alegria, I., Artola, X., Díaz de Ilarraza, A., Ezeiza, N. & Sologaitoa, A. 2007. ZT Corpus: Annotation and tools for Basque corpora. IN: M. Davies, P. Rayson, Hunston, S. & P. Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference*, University of Birmingham, UK.

- Arnold, D. 2003. Why translation is difficult for computers. IN: H. Somers (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins Publishing Company. pp.119-142.
- Arnold, D. J., Balkan, L., Meijer, S., Humphreys, R. Lee & Sadler, L. 1994. *Machine Translation: an Introductory Guide*. London: Blackwells-NCC.
- Atkins, S., Clear, J. & Ostler, N. 1992. Corpus design criteria. *Literary and Linguistic Computing*, 7(1), pp.1-16.
- Babych, B., Hartley, A. & Sharoff, S. 2009. Evaluation-guided pre-editing of source text: improving MT-tractability of light verb constructions. IN: *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, Barcelona, Spain. pp.36-43.
- Bäcklund, I. 1984. *Conjunction-Headed Abbreviated Clauses in English*. Acta universitatis Upsaliensis, Uppsala: Studia Anglistica Upsaliensia 50.
- Balkan, L., Netterz, K., Arnold, D. & Meijer, S. 1994. Test suites for natural language processing. IN: *Proceedings of Language Engineering Convention*, Paris, France. pp.17-22.
- Banerjee, S. & Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. IN: *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan. pp.65-72.
- Behrens, B. 1999. A dynamic semantic approach to translation assessment: ing-participial adjuncts and their translation into Norwegian. IN: M. Doherty (ed.) *Sprachspezifische Aspekte der Informationsverteilung*. Berlin: Akademie. pp.99-111.
- Belica, C. 1996. Analysis of temporal change in corpora. *International Journal of Corpus Linguistics*, 1(1), pp.61-74.
- Bernstein, T.M. 1981. *The Careful Writer - A Modern Guide to English Usage*. Atheneum.
- Bernth, A. 1997. EasyEnglish: A tool for improving document quality. IN: *Proceedings of the ACL 5th Conference on Applied Natural Language Processing*, Washington, DC.
- Bernth, A. 1998. EasyEnglish: Preprocessing for MT. IN: *Proceedings of the 2nd Controlled Language Applications Workshop (CLAW 1998)*, Pittsburgh. pp.30-41.
- Bernth, A. 1999a. Controlling input and output of MT for greater user acceptance. IN: *Proceedings of the 21th Conference of Translating and the Computer (ASLIB 1999)*, London.
- Bernth, A. 1999b. A confidence index for machine translation. IN: *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMIMT 1999)*, Chester, UK. pp.120-127.
- Bernth, A. 2006. EasyEnglishAnalyzer: taking controlled language from sentence to discourse level. IN: *Proceedings of the 5th International Workshop on Controlled Language Applications (CLAW 2006)*, Cambridge, Massachusetts.

- Bernth, A. & Gdaniec, C. 2000. *A translation confidence index for English-German MT*. IBM Research Report, Yorktown Heights, New York.
- Bernth, A. & Gdaniec, C. 2001. MTranslatability. *Machine Translation*. 16, pp175-218.
- Bernth, A. & McCord, M. 2000. The effect of source analysis on translation confidence. IN: J.S. White (ed.) *Proceedings of the Conference of the American Machine Translation Association (AMTA 2000)*, Springer-Verlag: Berlin & Heidelberg. pp.89-99.
- Biber, D. 1985. Investigating macroscopic textual variation through multifeature /multidimensional analysis. *Linguistics*, 23, pp.337-360.
- Biber, D. 1986. Spoken and written textual dimensions in English: resolving the contradictory finding, *Language*, 62, pp.384-414.
- Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. 1989. A typology of English texts, *Linguistics*, 27, pp.3-43.
- Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Biber, D., Conrad, S. & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S. & Leech, G. 2002. *Student Grammar of Spoken and Written English*. China: Longman.
- Blatz, J., Fitzgerald, E. Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A. & Ueffing, N. 2003. *Confidence estimation for statistical machine translation*. John Hopkins Summer Workshop Final Report.
- Bowker, L. & Pearson, J. 2002. *Working with Specialized Language. A practical guide to using corpora*. London/New York: Routledge.
- Brinton, L.J. 2000. *The Structure of Modern English: A Linguistic Introduction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Brown, P., Cocke, J., Della Pietra, S., Jelinek, F., Della Pietra, V., Lafferty, J., Mercer, R. & Rossin, P. 1990. A statistical approach to machine translation. *Computational Linguistics* 16, pp.79-85.
- Buck, T. 1999. *A Concise German Grammar*. Oxford/New York: Oxford University Press.
- Byrne, J. 2006. *Technical Translation. Usability Strategies for Translating Technical Documentation*. Netherlands: Springer.
- Cadwell, P. 2008. *Readability and Controlled Language*. Dublin City University. MA dissertation.
- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. 2008. Further meta-evaluation of machine translation. IN: *Proceedings of the 3rd Workshop on Statistical Machine Translation (WSMT 2008)*, Columbus, Ohio. pp.70-106.

- Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C. & Schroeder, J. 2007. (Meta-)evaluation of machine translation. IN: *Proceedings of the 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic. pp.136-158.
- Callison-Burch, C., Osborne, M. & Koehn, P. 2006 Re-evaluating the role of BLEU in machine translation research. IN: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy. pp.249-256.
- Calsamiglia, H. & Tusón, A. 1999. *Las cosas del decir: Manual del análisis del discurso*. Barcelona: Ariel.
- Carl, M. & Way, A. 2003. *Recent Advances in Example-based Machine Translation*. Dordrecht, Boston & London: Kluwer Academic Publishers.
- Carletta, J. 1996. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, 22(2), pp.249-254.
- Carpenter, W.T., James, I.K., Bilbe, G. & Bischoff, S. 2004. At issue: a model for academic/industry collaboration. *Schizophrenia Bulletin*, 30(4), pp.997-1004.
- Carter, R. & McCarthy, M. 2006. *Cambridge Grammar of English: A Comprehensive Guide. Spoken and Written English Grammar and Usage*. Cambridge University Press.
- Cavar, D., Küssner, U. & Tidhar, D. 2000. From Off-line Evaluation to On-line Selection. IN: W. Wahlster (ed.) *VerbMobil: Foundations of Speech-to-Speech Translation*. Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo: Springer.
- Charniak, E., Knight, K. & Yamada, K. 2003. Syntax-based language models for statistical machine translation. IN: *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX)*, New Orleans. pp.40-46.
- Cochran, W.G. 1963. *Sampling Theory*. 2nd ed. New York: John Wiley.
- Cormen, T., Leiserson, C., Rivest, R. & Stein, C. 2001. *Introduction to Algorithms*. 2nd ed. MIT Press.
- Coughlin, D. 2003. Correlating automated and human assessments of machine translation quality. IN: *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX)*, New Orleans. pp.23-27.
- CREST Working Group. 2008. Industry-led Competence Centres – Aligning academic / public research with Enterprise and industry needs. *Open Method of Co-ordination (OMC), 3rd Action Plan*.
- Dagan, I., Church, K.W. & Gale, W.A. 1993. Robust Bilingual Word Alignment for Machine Aided Translation. IN: *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Columbus, OH.
- De Preux, N. 2005. How much does controlled language improve machine translation results? IN: *Proceedings of the 27th Conference of Translating and the Computer (ASLIB 2005)*, London.

- DePalma, D.A. & Kelly, N. 2009. The business case for machine translation. How Organizations justify and adopt automated translation. Lowell, Massachusetts: Common Sense Advisory, Inc.
- Dervišević, D. & Steensland, H. 2005. *Controlled Languages in Software User Documentation*. University of Linköping. Master Thesis.
- Di Eugenio, B. & Glass, M. 2004. The Kappa statistic: a second look. *Computational Linguistics*, 30(1), pp.95-101.
- Díaz de Ilarraza, A., Labaka, G. & Sarasola, K. 2008. Statistical Post-editing: a valuable method in domain adaptation of RBMT systems for less-resourced languages. IN: *Proceedings of the Workshop Mixing Approaches to Machine Translation (MATMT 2008)*, Donostia, Spain.
- Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. IN: *Notebook Proceedings of the Human Language Technology*, San Diego. pp.128-132.
- Doherty, M. 1999. The grammatical perspective of -ing adverbials and their translation into German. IN: H. Hasselgård & S. Oksefjell (eds.) *Out of Corpora*. Amsterdam: Rodopi. pp.269-82.
- Duffley, P.J. 2005. *The English Gerund-Participle. A Comparison with the Infinitive*. Berkeley Insights in Linguistics and Semiotics, 61. New York: Peter Lang Publishing.
- Dugast, L., Senellart, J. and Koehn, P. 2007. Statistical post-editing on SYSTRAN's rule-based translation system. IN: *Proceedings of the 2nd Workshop on Statistical Machine Translation (WSMT 2007)*, Prague, Czech Republic. pp.220-223.
- Elming, J. 2006. Transformation-based correction of rule-based MT. IN: *Proceedings of the 11th Annual Conference of the European Association for Machine Translation (EAMT20 06)*, Oslo, Norway. pp.219-226.
- Emonds, J.E. 1985. *A Unified Theory of Syntactic Categories*. Dordrecht: Foris Publications.
- Espunya, A. 2007. Informativeness and explicit linking in the translation of the English V-ing free adjuncts into Catalan. *Language in Contrast*, 7(2), pp.143-166.
- Finch, A., Akiba, Y. & Sumita, E. 2004. How does automatic machine translation evaluation correlate with human scoring as the number of reference translations increases? IN: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal. pp.2019-2022.
- Font Llitjós, A., Carbonell, J. & Lavie, A. 2005. A framework for interactive and automatic refinement of transfer-based machine translation. IN: *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT 2005)*, Budapest, Hungary. pp.87-96.
- Fourth Workshop on Statistical Machine Translation. 2009. Greece, Athens. Proceedings available online at: <http://www.statmt.org/wmt09/WMT-09-2009.pdf>. [Last accessed on 31.10.09].

- Frederking, R. & Niremburg, S. 1994. Three heads are better than one. IN: *Proceedings of the 4th Conference on Applications for Natural Language Processing*, Stuttgart, Germany. pp.95-100.
- Frey, L.R., Botan, C.H., Friedman, P.G. & Kreps, G.L. 1991. *Investigating Communication. An Introduction to Research Methods*. London: Prentice Hall International.
- Gdaniec, C. 1994. The Logos translatability index. IN: *Proceedings of 1st Conference of the Association for Machine Translation in the Americas (AMTA 1994)*, Columbia, Maryland. pp.97-105.
- Gennrich, K. 1992. *Die Nachredaktion von Maschinenübersetzungen am Bildschirm – eine Prozeßuntersuchung*. Diploma thesis. University of Hildersheim.
- George, C. & Japkowicz, N. 2005. Automatic correction of French to English relative pronoun translations using natural language processing and machine learning techniques. IN: *3rd Computational Linguistics in the North-East Workshop (CLINE 2005)*, Ottawa, Canada.
- Giménez, J. & Marquèz, L. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. IN: *Proceedings of the ACL 2007 2nd Workshop on Statistical Machine Translation*, Prague, Czech Republic. pp.159-166.
- Greenhalgh, T. 1997. Statistics for the non-statistician, *BMJ*, 7104(315).
- Guzmán, R. 2007. Automating MT post-editing using regular expressions. *Multilingual*, 90, 18(6). pp.49-52.
- Guzmán, R. 2008. Advanced automatic MT post-editing. *Multilingual*, #95, 19(3). pp.52-57.
- Halliday, M.A.K. 2004. *An Introduction to Functional Grammar*. 3rd ed. London: Arnold.
- Hamon, O., Hartley, A., Popescu-Belis, A. & Choukri, K. 2007. Assessing human and automated quality judgments in the French MT evaluation campaign CESTA. IN: *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, Copenhagen, Denmark.
- Hargis, G. 2000. Readability and computer documentation. *ACM Journal of Computer Documentation*, 24(3). pp.122-131.
- Hashimoto, K., Yamamoto, H., Okuma, H., Sumita, E. & Tokuda, K. 2009. Reordering model using syntactic information of a source tree for statistical machine translation. IN: *Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, Boulder, Colorado. pp.69-77.
- Hauser, M.D., Chomsky, N. and Tecumseh Fitch, W. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298, pp.1569-1579.
- Heid, U. & Hildenbrand, E. 1991. Some practical experience with the use of test suites for the evaluation of SYSTRAN. IN: *Proceedings of the Evaluator's Forum*, Geneva, pp.195-213.

- Hochberg, J., Scovel, C. & Thomas, T. 1998. Bayesian stratified sampling to assess corpus utility. IN: E. Charniak (ed.), *Proceedings of the 6th Workshop on Very Large Corpora*, Montreal, Canada. pp.1-8.
- Hogg, M. & Voss, D.W. 2005 Same methods, different disciplines: the historian and linguist as technical communicators. IN: *Proceedings of the Conference of the Society for Technical Communication*, Seattle, Washington. pp.148-153.
- Huang, Fei & Papineni, K. 2007. Hierarchical system combination for machine translation. IN: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Prague, Czech Republic, pp. 277-286.
- Huddleston, R. & Pullum, G. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.
- Huijsen, W.O. 1998. Controlled language – An introduction. IN: *Proceedings of the 2nd Controlled Language Applications Workshop (CLAW 1998)*, Pittsburgh, Pennsylvania. pp1-15.
- Hunt, R.J. 1986. Percent agreement, Pearson's correlation, and Kappa as measures of inter-examiner reliability, *Journal of Dental Research*, 65(2), pp.128-130.
- Hurst, S. 2006. Authors vote for freedom of speech. *ISTC Newsletter*. May 2006. Institute of Scientific and Technical Communicators. Available at: http://www.istc.org.uk/Publications/Newsletter/News_2006/istcMay2006.pdf [Last accessed on 31.10.09].
- Hutchins, J., & Somers, H. 1992. *An Introduction to Machine Translation*. London: Academic Press Limited.
- Isabelle, P., Goutte, C. & Simard, M. 2007. Domain adaptation of MT systems through automatic post-editing. IN: *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, Copenhagen, Denmark. pp.255-261.
- Izquierdo, M. 2006. *Análisis Contrastivo y Traducción al Español de la Forma -ing Verbal Inglesa*. M.A. thesis. University of León, Spain.
- Izquierdo, M. 2008. *Contrastive Analysis and Descriptive Translation Study of English -ing Constructions and their Equivalents in Spanish*. PhD thesis. Universidad de León.
- Jaeger, P. 2004. Putting machine translation to work – Case study: language translation at Cisco. Presentation given at the LISA Global Strategies Summit. Foster City, California.
- Jakobson, R. 1959. On linguistic aspects of translation. IN: R. A. Brower (ed.) *On Translation*. Cambridge, MA: Harvard University Press/New York: Oxford University Press. pp.232-239.
- Jespersen, O. 1954. *A Modern English Grammar on Historical Principles*, Part II. London: George Allen & Unwin Ltd.

- Jurafsky, D. & Martin, J. 2000. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice Hall.
- Jurafsky, D. & Martin, J. 2009. *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. New Jersey/London: Pearson/Prentice Hall.
- Kamprath, C., Adolphson, E., Mitatumba, T. & Nyuberg, E. 1998. Controlled language for multilingual document production: Experience with Caterpillar Technical English. IN: *Proceedings of the 2nd Controlled Language Application Workshop (CLAW 1998)*, Pittsburgh, Pennsylvania. pp.51-61.
- Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London/New York: Longman.
- Kenney-Wallace, G. 2001. Partnerships, strategy and risk: research at the academic - industry interface. *ResearchResearch*. Available online at: <http://www.researchfortnight.co.uk/news.cfm?pagename=FundingArticle&ElementID=2882&lang=EN&type=Tech> [Last accessed on 31.10.09].
- King, M., Popescu-Belis, A. & Hovy, E. 2003. FEMTI: Creating and using a framework for MT evaluation. IN: *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX)*, New Orleans. pp.224-231.
- Knight, K. & Chander, I. 1994. Automated postediting of documents. IN: *Proceedings of the 12th National Conference of the American Association for Artificial Intelligence (AAAI 1994)*, Seattle, Washington, USA.
- Kochanski, G. 2006. *Statistical Sampling*. Available online at: <http://kochanski.org/gpk/teaching/0401Oxford/sampling.pdf> [Last accessed on 31.10.09].
- Koehn, P. & Monz, C. 2006. Manual and automatic evaluation of machine translation between European languages. IN: *NAACL Workshop on Statistical Machine Translation*, New York City, New York, USA. pp.102-121.
- Kohen, P., Hoan, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. IN: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics - Demo and Poster Sessions*, Prague, Czech Republic. pp.177-180.
- Kohl, J.R. 2008. *The Global English Style Guide: Writing Clear, Translatable Documentation for a Global market*. Cary, NC: SAS Institute Inc.
- Kortmann, B. 1991. *Free Adjuncts and Absolutes in English. Problems of Control and Interpretation*. London/New York: Routledge.
- Kretzschmar Jr., W., Darwin, C., Brown, C., Rubin, D. & Biber, D. 2004. Looking for the smoking gun: Principled sampling in creating the tobacco industry documents corpus, *Journal of English Linguistics*, 32(1), pp.31-47.

- Krings, H. P. 2001. *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, Ohio: The Kent State University Press. Edited/translated by G.S. Koby.
- Kulesza, A. & Shieber, S. 2004. A learning approach to improving sentence-level MT evaluation. IN: *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden.
- Lagarda, A.L., Alabau, V., Casacuberta, R., Silva, R. & Díaz de Ilarraza, E. 2009. Statistical post-editing of a rule-based machine translation system. IN: *Proceedings of the NAACL HLT 2009: Short Papers*, Boulder, Colorado. pp.217-220.
- Language Weaver News. 2008. Language Weaver CTO says improvements to statistical machine translation expand opportunities for government customers. July 9. Los Angeles.
- Laporte, P. 2009. SYSTRAN – Release 7. Presentation at Localization World Berlin, TAUS Workshop.
- Lassen, I. 2003. *Accessibility and Acceptability in Technical Manuals*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Lavie, A. Parlikar, A. & Ambati, V. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. IN: *Proceedings of the 2nd Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, Ohio. pp.87-95.
- LDC 2003. *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translation*. Project LDC2003T17.
- Lepage, Y. 1991. Test suites as a means for the evaluation of machine translation systems. IN: *Proceedings of the Evaluator's Forum*, Geneva, pp.225-235.
- Leech, G.N. 1987. *Meaning and the English Verb*. London/New York: Longman.
- Lehtola, A., Bounsaythip, C. & Tenni, J. 1998. Controlled language technology in multilingual user interfaces. IN: *Proceedings of the 4th ERCIM Workshop on User Interfaces for All (UI4ALL 1998)*, Stockholm, Sweden. pp.73-78.
- Leusch, G., Ueffing, N. & Ney, H. 2003. A novel string-to-string distance measure with applications to machine translation evaluation. IN: *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX)*, New Orleans. pp.240-247.
- Lin, C.Y. 2004. ROUGE: A package for automatic evaluation of summaries. IN: *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, Barcelona, Spain.
- Liu, D. & Gildea, C. 2005. Syntactic features for evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. IN: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, (ACL 2005)*, Ann Arbor, MI. pp.25-32.
- Loffler-Laurian, A.M. 1996. *La Traduction Automatique*. Lille: Presses Universitaires du Septentrion.

- Mailhac, J.P. 2000. Levels of speech and grammar when translating between English and French. IN: C. Shäffner & B. Adab (eds.) *Developing Translation Competence*. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp.33-50.
- Manning, C. & Schütze, H. 2000. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.
- McCaskill, M.K. 1998. Grammar, Punctuation, and Capitalization. A Handbook for Technical Writers and Editors. NASA SP-7084. Available online at [http://www.eknigu.org/get/L_Languages/LEn_English/McCaskill.%20Grammar,%20punctuation,%20and%20capitalization..%20a%20handbook%20for%20technical%20writers%20and%20editors%20\(NASA%20SP-7084\)\(108s\).pdf](http://www.eknigu.org/get/L_Languages/LEn_English/McCaskill.%20Grammar,%20punctuation,%20and%20capitalization..%20a%20handbook%20for%20technical%20writers%20and%20editors%20(NASA%20SP-7084)(108s).pdf) [Last accessed on 31.10.09].
- McEnery, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-based Language Studies*. London/New York: Routledge.
- Meadow, C.T., Boyce, B.R., Kraft, D.H. & Barry, C. 2000. *Text Information Retrieval Systems*. 2nd Ed. San Diego, California & London: Academic Press.
- Microsoft Corporation 1998. *Microsoft manual of style for technical publications*. 2nd. ed. Redmond, US: Microsoft Press.
- Moore, C. 2000. Controlled language at Diebold Incorporated. IN: *Proceeding of the 3rd International Workshop on Controlled Language Applications (CLAW 2000)*, Seattle, Washington. pp.51-61.
- Muegge, U. 2007. Controlled language: the next big thing in translation?, *ClientSide News Magazine*, 7(7), pp.21-24.
- Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. IN: A. Elithorn & R. Banerfi (eds.) *Artificial and Human Intelligence*, Amsterdam: North-Holland. pp.173-180.
- Nakazawa, T. & Kurohashi, T. 2009. Statistical phrase alignment model using dependency relation probability. IN: *Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, Boulder, Colorado. pp10-18.
- Neubert, 2000. Competence in language, in languages, and in translation. IN: C. Shäffner & B. Adab (eds.) *Developing Translation Competence*. Amsterdam/Philadelphia: John Benjamins Publishing Company. pp.3-18.
- Niehues, J., Herrmann, T., Kolss, M. & Waibel, A. 2009. The Universität Karlsruhe translation system for the EACL-WMT 2009. IN: *Proceeding of the 4th EACL Workshop on Statistical Machine Translation (WSMT 2009)*, Athens, Greece, pp.80-84.
- Nießen, S., Och, F.J., Leusch, G. & Ney, H. 2000. An evaluation tool for machine translation: Fast evaluation for MT research. IN: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece. pp.39-45.
- NIST Report 2002. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. On line at

<http://www.nist.gov/speech/tests/mt/2008/doc/ngram-study.pdf> [Last accessed on 31.10.09].

Nyberg, E., Mitamura, T. and Huijsen, W.O. 2003. Controlled language for authoring and translation. IN: H. Somers (ed.) *Computers and Translation: A Translator's Guide*. Amsterdam: John Benjamins Publishing Company. pp.245-281.

O'Brien, S. 2003. Controlling controlled English: an analysis of several controlled language rule sets. IN: *Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003)*, Dublin, Ireland. pp.105-114.

O'Brien, S. 2006. *Machine Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis*. PhD Thesis. Dublin City University.

O'Brien, S. Forthcoming. Controlled language and readability. IN: Angelone, E. a& G. Shreve (eds). *Translation and Cognition*. Amsterdam: John Benjamins.

O'Brien, S. and Roturier, J. 2007. How portable are controlled languages rules: a comparison of two empirical MT studies. IN: *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, Copenhagen, Denmark.

Och, F. J. 2003. Minimum error rate training for statistical machine translation. IN: *Proceedings of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z. & Radev, D. 2004. Syntax for Statistical Machine Translation. Final Report of Johns Hopkins 2003 Summer Workshop.

Och, F.J. & Ney, H. 2003a. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), pp.19-51.

Och, F.J. & Ney, H. 2003b. Minimum error rate training for statistical machine translation. IN: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan. pp.160-167.

Ogden, C. K. 1930. *Basic English: A General Introduction with Rules and Grammar*. London: Paul Treber & Co. Ltd.

Olive, J. 2005. *Global Autonomous Language Exploitation (GALE)*. DARPA/IPTO Proposer Information Pamphlet.

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. New York/Abingdon: Routledge.

Owczarzak, K. 2008. *A Novel Dependency-based Evaluation Metric for Machine Translation*. PhD thesis. Dublin City University.

Owczarzak, K., Graham, Y. & van Genabith, J. 2007a. Using f-structures in machine translation evaluation. IN: *Proceedings of the LFG07 Conference*, Stanford, CA. pp.383-396.

Owczarzak, K. van Genabith, J. & Way, A. 2007b. Dependency-based automatic evaluation for machine translation. IN: *Proceedings of the Workshop on Syntax and*

- Structure in Statistical Machine Translation (HLT-NAACL 2007)*, Rochester, NY. pp.86-93.
- Ozdowska, S. & Way, A. (2009) Optimal bilingual data for French-English PB-SMT. IN: *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation (EAMT 2009)*, Barcelona, Spain. pp.96-103.
- Papineni, K., Roukos, S., Ward, T. & Zhu, W. 2002a. BLEU: A method for automatic evaluation of machine translation. IN: *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL 2002)*, Philadelphia, Pennsylvania. pp.311-318.
- Papineni, K., Roukos, S., Ward, T., Henderson, J. & Reeder, F. 2002b. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French and Spanish results. IN: *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT 2002)*, San Francisco. pp.132-137.
- Pierce, J. (Chair) 1966. Language and Machines: computers in Translation and Linguistics. *Report by the Automatic Language Processing Advisory Committee (ALPAC)*. Publication 1416. National Academy of Sciences National Research Council.
- Popescu-Belis, A. 2003. An experiment in comparative evaluation: humans vs. computers. IN: *Proceedings of 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX)*, New Orleans.
- Przybocki, M., Sanders, G. & Le, A. 2006. Edit Distance: A metric for machine translation. IN: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. pp.2038-2043.
- Pym, J. F. 1988. Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system. IN: P. Mayorcás (ed.) *Proceedings of the 10th Conference of Translating and the Computer (ASLIB 1988)*, London: ASLIB. pp.80-96.
- Quah, C.K. 2006. *Translation and Technology*. New York: Palgrave Macmillan.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London & New York: Longman.
- Rabadán, R., Labrador, B. & Ramón, N. 2006. Putting meanings into words: English -ly adverbs in Spanish translation. IN: C. Mourón Figueroa & T.I. Moraleja Gárate (eds.) *Studies in Contrastive Linguistics*. Santiago de Compostela: Universidade de Santiago. pp.855-862.
- Rajman, M. & Hartley, A. 2001. Automatically predicting MT systems rankings compatible with fluency, adequacy or informativeness scores. IN: *Proceedings of the 4th ISLE Workshop on MT Evaluation, MT Summit VIII*, Santiago de Compostela, Spain. pp.29-34.
- Ramón García, N. 2003. *Estudio Contrastivo Inglés-español de la Caracterización de Sustantivos*. León: Universidad de León.
- Reppen, R., Fitzmaurice, S.M. & Biber, D. 2002. *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins Publishing Company.

- Reuther, U. 2003. Two in one – Can it work? Readability and translatability by means of controlled language. IN: *Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003)*, Dublin, Ireland. pp.124-132.
- Richards, W.D. 1998. *The Zen of Empirical Research*. Vancouver: Empirical Press.
- Rochford, H. 2005. *An Investigation into the Impact of Controlled Language across Different Machine Translation Systems: Translating controlled English into German*. MA thesis. Dublin City University.
- Römer, U. 2005. *Progressives, Patterns, Pedagogy. A Corpus-driven Approach to English Progressive Forms, Functions, Contexts and Didactics*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Roturier, J. 2004. Assessing a set of controlled language rules: Can they improve the performance of commercial machine translation systems? IN: *Proceedings of the 26th Conference of Translating and the Computer (ASLIB 2004)*, London. pp.1-14.
- Roturier, J. 2006. *An Investigation into the Impact of Controlled English Rules on the Comprehensibility, Usefulness and Acceptability of Machine-translated Technical Documentation for French and German Users*. PhD Thesis. Dublin City University.
- Roturier, J. 2009. Deploying novel MT technology to raise the bar for quality: A review of key advantages and challenges. IN: *Proceedings of the 12th Machine Translation Summit (MTS 2009)*, Ottawa, Canada.
- Roturier, J. & Lehmann, S. 2009. How to treat GUI option in IT technical texts for authoring and machine translation. *The Journal of Internationalisation and Localisation (JIAL)*, 1, pp.40-59.
- Roturier, J. & Senellart, J. 2008. Automation of post-editing in localization workflows. Presentation at LISA Europe Forum 2008, Dublin, Ireland.
- Roturier, J., Krämer, S. & Düchting, H. 2005. Machine Translation: The translator's choice. IN: *Proceedings of the 10th Localisation Research Centre Conference (LRC X)*, Limerick, Ireland.
- Russo-Lassner, G. Lin, J. & Resnik, P. 2005. A paraphrase-based approach to machine translation evaluation. Technical Report LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57, University of Maryland, College Park, MD.
- Ryan, J.P. 1988. The role of the translator in making an MT systems work: Perspective of a developer. IN: M. Vasconcellos (ed.) *Technology as Translation Strategy, American Translator Association Scholarly Monograph Series 2*, State University of New York at Binghamton SUNY. pp.127-132.
- Sågvall-Hein, A. 1997. Language control and machine translation. IN: *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*. Santa Fe, New Mexico.
- Sampson, G. 2002. *Empirical Linguistics*. London: Continuum.
- Schäfer, F. 2003. MT post-editing: How to shed light on the "unknown task": Experiences made at SAP. IN: *Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003)*, Dublin, Ireland. pp.133-140.

- Schiller, A. 2005. German compound analysis with wfsc. IN: *Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP 2005)*, Helsinki, Finland. pp.239-246.
- Schwenk, H., Abdul-Rauf, S., Barrault, L. & Senellart, J. 2009. SMT and SPE Machine Translation Systems for WMT'09. IN: *Proceedings of the 4th EACL Workshop on Statistical Machine Translation (WSMT 2009)*, Athens, Greece. pp.130-134.
- Schwitter, R., Ljungberg, A. and Hood, D. 2003. ECOLE: A look-ahead editor for a controlled language. IN: *Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003)*, Dublin, Ireland. pp.141-150.
- Senellart, J. 2007. SYSTRAN MT/TM Integration. *ClientSide News Magazine*, June 2007, Feature, pp.22-25.
- Shubert, S., Spyridakis, J., Holmback, H., and Coney, M. 1995. Testing the comprehensibility of Simplified English: An analysis of airplane procedure documents. *Journal of Technical Writing and Communication*, 25(4).
- Simard, M., Goutte, C. and Isabelle, P. 2007a. Statistical phrase-based post-editing. IN: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT 2007)*, Rochester, NY. pp.508-515.
- Simard, M., Ueffing, N., Isabelle, P. and Kuhn, R. 2007b. Rule-based translation with statistical phrase-based post-editing. IN: *Proceedings of the ACL 2nd Workshop on Statistical Machine Translation (WSMT 2007)*, Prague, Czech Republic. pp.203-231.
- Snoover, M., Dorr, B., Schwartz, R., Micciulla, L. & Makhoul, J. 2006. A study of translation edit rate with targeted human annotation. IN: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, Massachusetts.
- Somers, H. 1997. A practical approach to using machine translation software: "post-editing" the source text. *The Translator*, 3(2), pp.193-212.
- Somers, H. 2003. An overview of EBMT. In: Ed: Carl, M & Way, A. *Recent Advances in Example-based Machine Translation*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Spaggiari, L., Beaujard, F. and Cannesson, E. 2003. A controlled language at Airbus. IN: *Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003)*, Dublin, Ireland. pp.151-159.
- Spyridakis, J., Shubert, S. and Holmback, H. 1997. Measuring the translatability of Simplified English procedures. *IEEE Transactions on Professional Communication*, 40(1).
- SSST 2007. *First Workshop on Syntax and Structure in Statistical Translation (SSST-1)*, Rochester, New York.
- SSST 2008. *Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, Ohio.
- SSST 2009. *Third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, Boulder, Colorado.

- Surcin, S., Lange, E. & Senellart 2007. Rapid development of new language pairs at SYSTRAN. IN: *Proceedings of the 11th Machine Translation Summit of the International Association for Machine Translation (MT Summit XI)*, Copenhagen, Denmark.
- Tatsumi, M. & Sun, Y. 2008. A comparison of statistical post-editing on Chinese and Japanese. *Localisation Focus – The International Journal of Localisation*, 7(1), pp. 22-33.
- Thompson, H. 1991. Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment. IN: *Proceedings of the Evaluator's Forum*, Geneva, Switzerland. pp.215-223.
- Turcato, D., Popowich, F., McFetridge, P., Nicholson, D. & Toole, J. 2000. Preprocessing closed captions for machine translation. IN: C. Van Ess-Dykema, C. Voss and F. Reeder (eds.) *Proceedings of the NAACL-ANLP 2000 Workshop on Embedded Machine Translation Systems*, Seattle, Washington. pp.38-45.
- Turian, J., Shen, L. & Melamed, I.D. 2003. Evaluation of machine translation and its evaluation. IN: *Proceedings of the 9th Machine Translation Summit of the International Association for Machine Translation (MT Summit IX)*, New Orleans. pp.386-393.
- Tyers, F. & Donnelly, K. 2009. apertium-cy – a collaboratively-developed free RBMT system for Welsh to English. IN: E. Hajičová (ed.) *The Prague Bulletin of Mathematical Linguistics*. 91. pp.57-66.
- Underwood, N. & Jongejan, B. 2001. Translatability checker: A tool to help decide whether to use MT. IN: B. Maegaard (ed.) *Proceedings of the MT Summit VIII: Machine Translation in the Information Age (MTS VIII)*, Santiago de Compostela, Spain. pp.363-368.
- van Rijsbergen, C. 1979. *Information Retrieval*. 2nd ed. London: Butterworths.
- van Slype, G. 1979. *Critical methods for evaluating the quality of machine translation*. Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Technical Report BR-19142. Bureau Marcel van Dijk.
- Vasconcellos, M. 1987. Post-editing on-screen: machine translation from Spanish into English. IN: C. Picken (ed.) *Translating and the Computer 8: a profession on the move*. London: ASLIB. pp.133-146.
- Veale, T. & Way, A. 1997. Gaijin: A bootstrapping approach to example-based machine translation. IN: *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.
- Vertan, C. & v. Hahn, W. 2003. Menu choice translation – A flexible menu-based controlled natural language system –. IN: *Proceedings of the 4th Controlled Language Applications Workshop (CLAW 2003)*, Dublin, Ireland. pp.194-199.
- Viera, A.J. & Garrett, J.M. 2005. Understanding interobserver agreement: the Kappa statistic, *Family Medicine*, 37(5), pp.360-363.

- Vilar, D., Xu, J., D'Haro, L.F. & Ney, H. 2006. Error analysis of statistical machine translation output. IN: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.
- Vinay, J.P. & J. Darbelnet 1995. *Comparative Stylistics of French and English: a Methodology for Translation*. Translated by J. C. Sager and M. J. Hamel. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Way, A. & Gough, N. 2005. Controlled translation in an example-based environment: What do automatic evaluation metrics tell us? *Machine Translation*, 19, pp.1-36.
- Wells Akis, J. & Sisson, R. 2002. Improving translatability: A case study at Sun Microsystems, Inc. *The LISA Newsletter*. 11(4.5).
- Wendt, C. 2008. Large-scale deployment of statistical machine translation: Example Microsoft. Presentation given at: *The 8th Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, Waikiki, Hawai'i.
- White, F.D. 1996. *Communicating Technology: Dynamic and Models for Writers*. New York: HarperCollins College Publishers.
- White, J.S., O'Connell, T. & O'Mara, F. 1994. The ARPA MT evaluation methodologies: Evolution, lessons and further approaches. IN: *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*. Columbia, MD. pp.193-205.
- Wierzbicka, A. 1988. *The Semantics of Grammar*. Amsterdam/Philadelphia: John Benjamins Publishing Company.
- Wojcik, R.H., Hoard, J.E. & Holzhauser, K.C. 1990. The Boeing Simplified English Checker. IN: *Proceedings of the International Conference on Human-Machine Interaction and Artificial Intelligence in Aeronautics and Space*, Toulouse, France. pp.43-57.
- WSMT 2009. *4th EACL Workshop on Statistical Machine Translation*, Athens, Greece.
- Zaidan, O. & Callison-Burch, C. 2009. Feasibility of human-in-the-loop minimum error rate training. IN: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Singapore. pp.52-61.
- Zollmann, A., Venugopal, A., Och, F. & Ponte, J. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. IN: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK. pp.1145-1152.

APPENDICES

APPENDIX A: COMPLETE FUNCTIONAL CLASSIFICATION OF THE –ING WORDS IN THE CORPUS

Titles	Pattern	Position1	Position2	№ of examples
	-ing	Free		1,255
		Embedded	Embedded	620
	About + -ing		Beginning of sentence	530
		Free		100
		Embedded	Beginning of sentence	60
Characterisers			Embedded	38
	Position	Type		№ of examples
	Pre-modifiers	modifiers (participial adjectives and gerundial nouns)		1,873
	Post-modifiers	reduced relative clauses		377
		nominal adjuncts		226
		adjectival adjuncts		12
Adverbials	Clause type	Prep. or sub. conj.		№ of examples
	Manner	By		516
		Free		159
		Without		88
	Time	When		313
		Before		179
		After		139
		While		65
		On		8
		Through		5
		Between		4
		Free		3
		Along with		2
		During		2
		From		2
		In		2
		In the middle of		2
		Prior		1
		Upon		1
	Purpose	For		443
		In		1
	Condition	If		20
	Contrast	Instead of		11
	Cause	Because		2
	Concession	Besides		1
	Place	Where		1
Progressives	Time	Voice		№ examples
	Present	Active		501
		Passive		117
		Questions		2
	Modal			22
	Past	Active		9
		Passive		3
	Infinitive			5
	Future			2
Referentials	Types			№ examples
	Nouns			252
	Catenative verbs			167
	Prepositional verbs			116
	Comparatives			46
	Phrasal verbs			13

APPENDIX B: ING EVALUATION GUIDELINES

The aim of this evaluation is to analyze the machine translation of ing words to gather information about the efficiency with which our rule-based machine translation (MT) system handles these constituents. For this purpose, an excel file is provided with a layout for evaluation. The file contains sentences which include ing words. They were extracted from Symantec user guides for three different products. These have been machine translated for evaluation.

The question you are asked to answer is the following:

- Is the machine translation of the ing word grammatical and accurate?

If both requirements are true for your target language, the example should be considered "Correct". If any of the requirements is not true, the example should be considered "Incorrect". Please note that only the constituent of study should be taken into account, not the entire sentence. Type an "X" in the appropriate cell in the excel file provided.

Note that in addition to the source sentences and the raw MT output you are asked to evaluate, a post-edited (PE) version has been provided. A language offers many ways of expressing the same idea. The PE version is one that our post-editors thought met the Symantec standards. You may disagree with the PE version and consider the MT output correct or incorrect. We do not want you to check whether the raw MT output and PE version are the same and classify accordingly. That could be done with a computer!! Remember, however, that we are not aiming for a perfect output but for a version that is grammatical and accurate.

A number of issues that English ing words can pose for rule-based MT systems are described below. They should be taken into account when deciding whether the translation of the ing word is correct or not. Should the ing words pose any other problem for your particular target language, these should also be considered.

1-Gerundial nouns, gerund-participles or participial adjectives?

Rule-based MT systems carry out a source analysis step where they assign the correspondent grammatical category to each of the words in the sentence (tagging). For doing so, they rely heavily on the words following and preceding the word to be

tagged. Ing words can be gerundial nouns (act as genuine nouns), participial adjectives (act as adjectives) and gerund-participles (have verbal and noun/adjective characteristics). This flexibility of the English language poses problems for MT rule-based MT systems because the different types of ing words can appear in the exact environment, that is, followed and preceded by the same word category. This means that MT systems do not have enough information for discrimination and they have to “guess” the category of the ing words.

- Backup Exec includes a diagnostic application (Bediag.exe) that gathers information about a Windows XP, Windows 2000, or Windows Server 2003 computer for **[troubleshooting]** purposes.]
- " Checklist **[for troubleshooting]** [devices that have gone offline] " on page 1544

The previous sentences contain the phrase “for troubleshooting” followed by the nouns “purposes” and “devices” respectively. Even if the words surrounding the ing word belong to the same category, we observe that the first is a participial adjective and the second a gerund-participle. Depending on the target language, the ing word might need to be translated using two different structures. Hence, the need for MT systems to discriminate between the different categories.

- In the Specify SSH utility and required parameters **[for configuring]** [UNIX Targets dialog box], you can enter a command line for connecting to the UNIX target machine.
- Back up files and directories **[by following]** [junction points]
- To create an XML file **containing** all parameters, use the /XML:
- The name of the computer **running** the remote agent.

Similarly, the highlighted ing words in the examples above are gerund-participles. However, the position in which they appear could be filled by either a gerund-participle or a participial adjective or even a gerundial noun, the MT system might have difficulties for discrimination. This, in turn, could result in an incorrect translation of the ing word.

Check these issues when evaluating the ing form. In order to facilitate the identification of these structures, the ing word for evaluation has been highlighted in each sentence.

2-Participial adjectives:

Ing words can act as adjectives (participial adjectives). This is the case of “existing” and “error-handling” below. In some languages, adjectives must agree in number and gender with the noun they are modifying. In others, specific connectors are required to bind adjectives to the noun they are modifying. In addition, they sometimes need to appear in specific positions in the sentence.

- Check No loss restore (do not delete existing transaction logs) to preserve [the **existing** transaction logs] on the Exchange 5.5 Server.
- To apply an error-handling rule for a specific error code that is in an error category, you can create [a custom **error-handling** rule.]

Check these requirements are met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, the ing word for evaluation has been highlighted in each sentence.

3- INGs and implicit subjects

Ing words are often heads of non-finite clauses, which allow for their subject to be implicit. According to the English grammar, the subject of a non-finite clause can only be omitted if the subject of the non-finite clause and the subject of the main clause are the same.

- **After running** the database snapshot job, Backup Exec creates history and job log information to indicate the job's status.
 - After Backup Exec/you/? runs the database snapshot job, Backup Exec creates history and job log information to indicate the job's status.
- **Before implementing** the IDR option in a CASO environment, review the following:
 - Before you/? implement the IDR option in a CASO environment, review the following:

However, this requirement is not always met and it is not unusual to find sentences like the following:

- This option is only available **when performing** full backups.
 - This option is only available when this option/you/? performs full backups.
- Displays information detailing what has occurred **while running** the Command Line Applet and the specified option.
 - Displays information detailing what has occurred while it/you/? runs the Command Line Applet and the specified option.

Rule-based MT systems are programmed to follow grammatical structures. Therefore, if required to “recover” the subject for the subordinate clause, following the grammatical rule explained above, they may look for the subject of the main clause and assign the same to the subordinate clause. However, if we look at the examples above, this rule is not met. For instance, it is clear that “this option” is not what “performs full backups”. It could be argued that the subject of the subordinate clause is “you” or “the machine/programme”. As a consequence, if the target language requires knowing the subject of the subordinate clause to render a correct translation, this will be assigned incorrectly and the translation will be incorrect.

Check these requirements are met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, subordinate conjunctions and the ing words have been highlighted.

4. ING as the required form

A word might be required to appear in its ing form by the preceding verb or noun/adjective + preposition structure. It is sometimes difficult for MT systems to guess the right form for the target language.

- However, you might **consider adding** the media server name to the report name.
- Symantec **recommends using** fully qualified computer names.
- Select this option to **prevent** Backup Exec **from overwriting** files on the target disk with files that have the same names that are included in the restore job.
- A suggested resolution to **assist in recovering** from the error in a timely manner.
- Refer to your Microsoft Windows documentation for **information on creating** a Group Policy Object.
- **Requirements for creating** a removable backup-to-disk folder " on page 232
- We hope that you have found this document **useful in learning** more about the product.

Check these requirements are met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, the verbs and nouns/adjectives + prepositions have been highlighted as well as the ing words.

5. ING for progressive aspect

In English, the progressive aspect of a verb is marked by using its ing form. This might be done in a different way in other languages, where adverbs or particular declinations might need to be introduced to convey aspectual information.

- On the server that hosts the RSG, there must be a storage group with the same name as the original storage group for the data **you are restoring**.
- If **you are using** a tape device that is not serialized, you need to manually eject the media from the device or reboot the device
- If this media is password protected and **is being cataloged** by this system for the first time, enter the password.
- If failover occurs in the middle of backing up a resource, the media that **was being used** at the time of the failover is left unappendable and new media will be requested upon restart.

Check these requirements are met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, the verbal periphrasis has been highlighted in each sentence.

6. INGs in titles

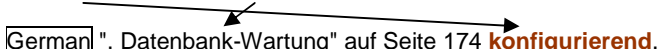
Titles can be non-finite clauses with ing words as heads. This particular structure might require a specific translation which does not conform to the usual translation of ing words.

- **Installing** servers for replication
- **Configuring** SMTP email or mobile phone text messaging for a person recipient
- **About performing** a DBA-initiated backup job for Oracle
- **"Creating** selection lists" on page 340
- For more information about file exclusion, see **"Including or excluding** files for backup" on page 337 .

Check this requirement is met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, the head ing words have been highlighted for titles.

7. Position

When translating an ing word, the MT system might need to reorder the words of the target sentence and this might not always be done correctly. The following example presents two problems. First of all, "Configuring" has been incorrectly recognized as a participial adjective. This means that the translation will be incorrect (see section 1). In addition, participial adjectives should be placed in front of the noun they modify, "Datenbank-Wartung" in this case. This requirement has not been met, as the translation for Configuring has been placed at the end of the sentence, after the page number.

- **"Configuring** database maintenance" on page 174.

German: ", Datenbank-Wartung" auf Seite 174 **konfigurierend**.

Check these requirements are met when evaluating the ing form. In order to facilitate the identification of these structures, the ing word to be evaluated has been highlighted in each sentence.

8. Transposition

When translating an ing word, the MT system might resource to transposition. For instance, in the example below, the ing word has been translated as a noun into Spanish. The adverb has been maintained. This results in a noun being modified by an adverb, which is incorrect. Nouns must be modified by adjectives. This example,

therefore, is incorrect because the MT system was not able to handle an ing word preceded by an adverb.

- A person [...] should be charged with constantly **supervising** your organizational disaster preparation efforts.

preposition adverb gerund-participle

Spanish: Una persona debe ser cargada con constantemente la **supervisión** de sus esfuerzos de la preparación del desastre organizacional.

prep. adverb noun

Check these requirements are met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, the ing word to be evaluated has been highlighted in each sentence.

9. Attachments

In some cases, other pieces of the sentence need to be embedded or added to the translation of an ing word. For instance, in the example below, the pronoun “it” needs to be attached to the translation of the ing word. It is therefore necessary to make sure that the pronoun attached to the ing word is in the right form (gender and number in the case of Spanish). This is an example to check whether the MT system is able to handle ing words that requires attachments.

- Confirm the password **by re-entering** it in the Confirm sa Password field.

structure pronoun

Spanish: Confirme la palabra de paso **volviéndola a entrar** en el campo de la palabra de paso confir

singular feminine noun gerund singular verb feminine pronoun

Check these requirements are met for your target language when evaluating the ing form. In order to facilitate the identification of these structures, the ing words have been highlighted.

10. No ING error

Not all problems that occur in the environment of the ing are due to the MT system not been able to handle the ing word correctly. For instance, it might be the case where the preposition is translated incorrectly (see example below). This should not be considered as an incorrect translation of the ing word.

- Creating a restore job **while reviewing** media or devices

preposition

Spanish: Crear un trabajo del restore **mientras que repasa** media o los dispositivos

preposition

Make sure that these issues do not affect the evaluation of the ing form.

Layout of the excel file⁶⁰:

- Column A "source sentences" contains English sentences that include ing words. Note that the constituent of study is highlighted in red to help identify it easily.
- Column B "raw MT output" contains the machine translated output of the source sentences. The source sentences were machine translated using the red highlighting and the MT preserved it where it thought it belonged in the target language. Please note that the highlight was kept only to help identify the constituents. It is not always accurate or even present in the target language examples.
- Column C "post-edited version" contains a post-edited version of the raw MT output. This version has been provided to as a possible final version. You may disagree with the PE version and consider the MT output correct or incorrect. But remember that we are not aiming for a perfect output but for a version that is grammatical and accurate.
- Use columns D "correct" and E "incorrect" to answer the question. Please type an "x" in the appropriate cell.

	A	B	C	D	E
1	STUDY DATA FIELDS				
2	SOURCE SENTENCE	RAW MT OUTPUT	POST-EDITED VERSION	ANSWER FIELDS	
3	After you install the server used for failover or load balancing , you need to configure it with the Symantec Endpoint Security Management console.	Una vez que instale al servidor utilizado para la conmutación por error o balanceo de carga, usted necesita configurarlo con la consola de administración de Symantec Endpoint Security.	Una vez instale el servidor utilizado para la conmutación por error o balanceo de carga, necesita configurarlo con la consola de administración de Symantec Endpoint Security.	CORRECT	INCORRECT
4	Restoring a system backed up with a working set strategy requires only the media containing the latest working set backup media and the media containing the most recent full backup.	Restaurando un sistema hecho copia de respaldo de con una estrategia del espacio de ejecución necesita solamente los soportes que contienen los últimos soportes de copia de respaldo del espacio de ejecución y los soportes que contienen la copia de respaldo completa más reciente.	La restauración de un sistema del que se ha realizado una copia de respaldo con una estrategia de espacio de ejecución necesita solamente los soportes que contienen la última copia de respaldo del espacio de ejecución y los soportes que contienen la copia de respaldo completa más reciente.		
5	A suite of pre-defined queries assist you in identifying key issues such as database integrity, security, and permissions tracking.	Un conjunto de consultas predefinidas le asiste en identificar las cuestiones claves tales como integridad de base de datos, seguridad, y seguimiento de los permisos.	Un conjunto de consultas predefinidas le asiste en la identificación de cuestiones claves tales como integridad de base de datos, seguridad, y seguimiento de los permisos.		
6	The SANS Top 10 Policy is a set of procedures for securing a machine	SIN la política de la tapa 10 es un conjunto de procedimientos para asegurar un equipo	La norma SANS Top 10 es un conjunto de procedimientos para asegurar un equipo		

Before submitting, please check that an answer has been provided for all the rows.

⁶⁰ Please write your answers in the columns assigned for this purpose and do not modify the cells with content.

APPENDIX C: QUESTIONNAIRE FOR EVALUATORS

1- Are you a full-time or a part-time translator?

☐ Full-time

☐ Part-time

2- Please specify the rough number of words you have translated.

3- Have you ever taken courses on your native language grammar?

☐ At university

☐ In high-school

☐ No

4- Have you ever post-edited machine translation output professionally?

☐ Yes

☐ No

5- If you answered yes to question number 4, please specify the rough number of words you have post-edited.

6- Do you like working with machine translation output?

☐ 1- Not at all

☐ 2- Somewhat

☐ 3- Moderate

☐ 4- Very much

APPENDIX D: LIST OF ABBREVIATIONS FOR -ING SUBCATEGORIES

Adv_C_because	Adverbial of cause introduced by "because"
Adv_CC_besides	Adverbial of concession introduced by "besides"
Adv_CD_if	Adverbial of condition introduced by "if"
Adv_CT_insteadof	Adverbial of contrast introduced by "instead of"
Adv_E_0	Adverbial of elaboration with no introductory preposition/subordinate conjunction
Adv_M_0	Adverbial of mode with no introductory preposition/subordinate conjunction
Adv_M_by	Adverbial of mode introduced by "by"
Adv_M_with	Adverbial of mode introduced by "with"
Adv_M_without	Adverbial of mode introduced by "without"
Adv_PL_where	Adverbial of place introduced by "where"
Adv_PU_0	Adverbial of purpose with no introductory preposition/subordinate conjunction
Adv_PU_for	Adverbial of purpose introduced by "for"
Adv_PU_in	Adverbial of purpose introduced by "in"
Adv_R_0	Adverbial of result with no introductory preposition/subordinate conjunction
Adv_T_0	Adverbial of time with no introductory preposition/subordinate conjunction
Adv_T_after	Adverbial of time introduced by "after"
Adv_T_alongwith	Adverbial of time introduced by "along with"
Adv_T_before	Adverbial of time introduced by "before"
Adv_T_between	Adverbial of time introduced by "between"
Adv_T_during	Adverbial of time introduced by "during"
Adv_T_from	Adverbial of time introduced by "from"
Adv_T_in	Adverbial of time introduced by "in"
Adv_T_inthemiddleof	Adverbial of time introduced by "in the middle of"
Adv_T_on	Adverbial of time introduced by "on"
Adv_T_prior	Adverbial of time introduced by "in"
Adv_T_through	Adverbial of time introduced by "through"
Adv_T_upon	Adverbial of time introduced by "upon"
Adv_T_when	Adverbial of time introduced by "when"

Adv_T_while	Adverbial of time introduced by “while”
Char_POadj	Characteriser post-modifier adjectival adjuncts
Char_POnn	Characteriser post-modifier nominal adjuncts
Char_POrr	Characteriser post-modifier reduced relative clauses
Char_PR	Characteriser pre-modifier
Prog_fut	Progressive future tense
Prog_inf	Progressive in the infinitive form
Prog_mod	Progressive with a modal auxiliary
Prog_PRact	Progressive present tense active voice
Prog_PRpas	Progressive present tense passive voice
Prog_PRq	Progressive present tense question
Prog_PSact	Progressive past tense active voice
Prog_PSpas	Progressive past tense passive voice
Ref_cat	Referential catenative verb
Ref_comp	Referential in comparative structure
Ref_nn	Referential noun
Ref_phrV	Referential phrasal verb
Ref_prepV	Referential prepositional verb
Title_ABING_indp	Title starting with “about + ing” independent
Title_ABING_QM1	Title starting with “about + ing” embedded within quotation marks at the beginning of sentence
Title_ABING_QM2	Title starting with “about + ing” Embedded within quotation marks Embedded in sentence
Title_ING_indp	Title starting with “ing” independent
Title_ING_QM1	Title starting with “ing” embedded within quotation marks at the beginning of sentence
Title_ING_QM2	Title starting with “ing” embedded within quotation marks Embedded in sentence

APPENDIX E: GUIDELINES FOR THE ANALYSIS OF GERMAN AND JAPANESE -ING EVALUATION

The aim of this analysis is to pinpoint the errors generated by our rule-based machine translation (MT) system when it deals with -ing constituents. In a previous stage, an evaluation was carried out where professional native language translators judged whether the raw machine translation of -ing constituents were translated grammatically and accurately. The former is concerned with following the rules governing the use of a language and the latter considers whether the meaning of the original text was preserved.

The question you are now asked to answer is the following:

Why is the machine translation of the -ing constituent incorrect?

An excel file is provided with a layout for analysis. It contains the source sentences and their raw MT output, together with a post-edited (PE) version. A language offers many ways of expressing the same idea. The PE version is one that our post-editors thought met the Symantec standards. Please use this version only as a guideline and do not base your analysis on the comparison between the raw MT output and the PE version. All the sentences in the file were classified as incorrect by evaluators.

Moreover, based on the analysis of other target languages, a list of possible errors was compiled and is provided (see detailed information below). If the problem in your target language matches any of the errors in the list, type an X in the corresponding column. If you cannot find a match, please describe the problem in the "Other" column.

I am available for consultation in case of any doubt. Please email your question to nora.aranberrimonasterio@dcu.ie

TERMINOLOGY

We observed that some instances were classified as incorrect by evaluators due to terminological inaccuracy or because terms were left untranslated by the MT system, for example:

EN: A third party ISO 9660-compliant CD **burning** application to burn the IDR-created bootable CD image to a CD.
FR MT: Une application **brûlante** CD conforme d'OIN 9660 de tiers pour brûler l'image Différence-crée de CD bootable à un CD.
Correct terminology: Une application tierce **de gravure** de CD

EN: Enter a new name or change the name for the **error-handling** rule.
ES MT: Escriba un nuevo nombre o modifique el nombre para la norma de la error-**administración**.
Correct terminology: norma **de administración de errores**

GRAMMAR

Modifier instead of gerund-participle. The -ing word was recognized as a modifier instead of a gerund-participle. That is, the -ing word was recognized as an item modifying a noun (backing up database = a database used to back up something) rather than as an item referring to an action (backing up database = the system is backing up a database). See examples below:

EN: **Backing up** database files in a Microsoft cluster
FR MT: Bases de données **de sauvegarde** dans une batterie de Microsoft
Gloss: Databases **for backing up** in a battery of Microsoft
Correct version: **Sauvegarder** des bases de données dans un cluster Microsoft

EN: **Viewing** SCSI information for a robotic library" on page 225
ES MT: "Información del SCSI **de la visualización** para una biblioteca robótica" en la página 225
Gloss: "SCSI information **of the viewing** for a robotic library" on page 225
Correct version: "**Ver** información SCSI para una biblioteca robótica" en la página 225

Agreement mistake. The number/gender agreement required between head-modifier failed. This could also happen to subject-verb agreements, declination agreements, etc. If any of these errors is present, please specify the agreement mistake. See examples below:

EN: Before reverting a database, Backup Exec deletes all **existing** database snapshots, including those created with SQL 2005, with the exception of the snapshot used for the revert.

FR MT: Avant de retourner une base de données, Backup Exec efface tous les instantanés de base de données **existante**, y compris ceux créés avec SQL 2005, excepté l’instantané utilisé pour le retour.

Explanation: *instantanés* is a plural masculine noun whereas the adjective *existante*, which is modifying it, is feminine singular.

Correct version: Avant de réinitialiser une base de données, Backup Exec supprime tous **les clichés de base de données existants**, y compris ceux créés avec SQL 2005, excepté le cliché utilisé pour la réinitialisation.

EN: Note the **following** items when using continuous protection as part of your backup strategy:

FR MT: Tenga en cuenta **lo siguiente elementos** al usar la protección continua como parte de su estrategia de copia de respaldo:

Explanation: *elementos* is a plural noun but the adjective *siguiente*, which is modifying it, is in singular form.

Correct version: Tenga en cuenta **los siguientes elementos** al usar la protección continua como parte de su estrategia de copia de respaldo:

Progressive aspect loss. When translating from English into a target language, the progressive aspect introduced by the –ing word was often not kept. This might be correct or incorrect depending on the target language requirements. See examples below:

EN: If failover occurs in the middle of backing up a resource, the media that **was being used** at the time of the failover is left unappendable and new media will be requested upon restart.

ES MT: Si la conmutación por error ocurre en el medio de hacer copia de respaldo de un recurso, los soportes que **era utilizado** a la hora de la conmutación por error se dejan los soportes unappendable y nuevos serán solicitados sobre reinicio.

Explanation: The Spanish equivalent of “was used” was generated by the MT system instead of “was being used”. The imperfect preterite (tense referring to an action happening in a time previous to the moment of speaking) does not focus on the fact that the media was being used in the precise moment in which the backing up was happening. The tense used is more general, referring to all the times where the backing up happened in the past.

Correct version: Si la conmutación por error ocurre durante el proceso de copia de respaldo de un recurso, los soportes que **estaban siendo utilizados** a la hora de la conmutación por error se dejan unappendable y se solicitan soportes nuevos durante el reinicio.

EN: In a cluster environment, if a job created through BACKINT or RMAN **is processing** and the node fails over, the job operation does not restart from the point when the node went down.

FR MT: Dans un environnement de batterie, si un emploi créé par BACKINT ou RMAN **traite** et l’échouer de noeud plus de, l’exécution du travail ne relance pas du point quand le noeud est descendu.

Explanation: The French equivalent of “processes” was used instead of “is processing”. The present continuous tense was translated by a present simple tense, which refers to how actions happen in general and not in a precise moment of speaking.

Correct version: Dans un environnement de cluster, si un emploi créé par BACKINT ou RMAN **est en cours de traitement** et si le nœud bascule, l’opération de travail ne redémarre pas du point auquel le nœud a basculé.

Incorrect word class. The word class into which the -ing word was translated is incorrect. For instance, it was translated into a gerund (creando) when it should have been translated into an infinitive (crear). Please specify the word class into which the -ing word was translated and it should have been translated into. See examples below:

EN: Go to “Restoring data by setting job properties” on page 498 for **details on selecting** restore job properties.

FR MT: Allez au “Restaurant des données en plaçant des propriétés du travail” à la page 498 pour des **détails sur choisir** des propriétés du travail de restauration.

Explanation: “détails sur choisir”, where the -ing word is translated into an infinitive, was generated by the MT instead of “détails sur le choix”, where the -ing word is translated into a noun.

Correct version: Allez à “Restaurer des données en définissant des propriétés de travail” à la page 498 pour des **détails sur la sélection** des propriétés du travail de restauration.

ES: **Pausing** and resuming devices

ES MT: **Interrumpiendo** momentáneamente y continuando los dispositivos

Explanation: The gerund form “interrumpiendo” was used instead of the infinitive “interrumpir”.

Correct version: **Interrumpir** momentáneamente y reanudar los dispositivos

Incorrect preposition/subordinate conjunction. The preposition or subordinate conjunction preceding the -ing word were incorrectly translated. See examples below:

EN: If this attempt fails, CASO makes a second **attempt at finding** another available managed media server to process the jobs.

FR MT: Si cette tentative échoue, CASO fait un deuxième **tentative à trouver** un autre serveur multimédia contrôlé disponible pour traiter les travaux.

Explanation: The structure “tentative à” was generated by the MT system instead of the correct “tentative de”, which is, in turn, followed by a noun.

Correct version: Si cette tentative échoue, CASO fait une deuxième **tentative de détection** d’un autre serveur de supports géré disponible pour traiter les travaux.

ES: **While using** DLO, you can suppress dialogs by checking the Don't show me this check box message again.

ES MT: **Mientras que usa el** DLO, usted puede suprimir cuadros de diálogo activando no me muestra esta casilla de selección del mensaje de nuevo.

Explanation: The MT system generated the subordinate conjunction “mientras que”, closer to the meaning of “whereas” and conveying a sense of opposition, instead of “mientras”, which refers to simultaneous events, as “while” does.

Correct version: Mientras usa el DLO, puede suprimir cuadros de diálogo activando “No mostrar de nuevo este mensaje de casilla de selección”.

Incorrect head modified. The –ing word modified the incorrect head, for example:

EN: “Copy database” on page 55 lets you copy an **existing** Backup Exec Database from a computer using a specific name to another computer using the same name.
FR MT: Le "Copiez la base de données" à la page 55 vous laisse copier un Backup Exec **existant** Database à partir d'un ordinateur utilisant le même nom.
Explanation: The French adjective “existant” was inserted in the middle of the noun “Backup Exec Database” and is therefore only modifying part of the head it should be modifying.
Correct version: “Copier la base de données” à la page 55 vous permet de copier une base de données Backup Exec **existante** à partir d'un ordinateur utilisant un nom spécifique vers un autre ordinateur utilisant le même nom.

EN: Support for the DB2 log **archiving** methods that are known as user exit and VENDOR.
ES MT: Ayuda para el registro DB2 **que archiva los** métodos que son conocidos como la salida de usuario y DISTRIBUIDOR.
Explanation: “archiving” is modifying the noun “methods” and the complex noun “archiving methods” is modified by “DB2 log”. Therefore the translation should be "métodos de archivado para registros DB2" and not “registro DB2 que archiva los métodos” (DB2 log which archives the methods).
Correct version: Ayuda para los métodos de archivado de registros DB2 conocidos como la salida de usuario y DISTRIBUIDOR.

Inaccurate grammatical structure. The grammatical structure used, although valid, did not convey the same meaning as the source text, making the sentence inaccurate, for example:

EN: dr file contains specific information for the computer being protected, including:
FR MT: DR. file contient l'information spécifique pour **l'ordinateur étant protégé**, incluant :
Explanation: The structure “l'ordinateur étant protégé ” is grammatical. However, instead of meaning "the computer being protected" acquires de meaning of "once the computer is protected", which does not convey the same meaning as the source structure.
Correct version: Le fichier dr contient des informations spécifiques pour **l'ordinateur protégé**, incluant :

Misuse of pronouns. Pronouns were inserted where they were not required and vice versa. Please mark any problems with any type of pronouns in this column and specify the exact issue. See examples below:

EN: From the Administration Console, ensure you are connected to the Primary Group Server and that you **are running** in partition management mode.
ES MT: De la consola de administración, asegúrese que le conecten al servidor primario del grupo y eso que usted **se está ejecutando** en modo de la administración de la partición
Explanation: The Spanish translation was generated using the reflexive pronoun "se", which is not correct as it is not the user who has to be run (you are running yourself in partition management mode), but the program that has to be run by the user (you are running [the computer] in partition mode).

Correct version: Desde la consola de administración, asegúrese de que está conectado al servidor de grupo primario y de que está ejecutando en modo de partición.

Ungrammatical passive structure. Instances were found where passive structures were not correctly constructed. Please note that if the structure is grammatical but cumbersome, this should be marked in the "cumbersome structures" column instead of here. See examples below:

EN: Consider what type of Windows computer **is being protected**, the available hardware, and the system BIOS when selecting the type of bootable media to create.

ES MT: Considere qué tipo de equipo de Windows **está estando protegido**, el hardware disponible, y el sistema BIOS al seleccionar el tipo de soportes de arranque para crear.

Explanation: The translation “está estando protegido” is ungrammatical as passive structures in Spanish are constructed using the auxiliary “ser” and not “estar” (both translated as to be in English). The correct translation would have been “está siendo protegido”.

Correct version: Considere qué tipo de equipo de Windows está siendo protegido, el hardware disponible y el sistema BIOS, al seleccionar el tipo de soportes de arranque que se deben crear.

Ungrammatical implicit subject. According to grammar, the implicit subject of a subordinate clause must be the same as the subject of the main clause. It was observed that some examples did not comply with this rule. See examples below:

EN: This option is only available **when performing** full backups.

FR MT: Cette option est seulement disponible **en exécutant de** pleines sauvegardes.

Explanation: The subject of the subordinate clause “when performing full backups” is not “this option” but “you”. Therefore, according to English grammar, the subject of the subordinate clause “you” should be made explicit. This results in an incorrect translation because the MT system was expecting a grammatical sentence, or in the translation in the target language also being ungrammatical because the same grammatical rule applies.

Correct version: Cette option est seulement disponible **lors de l'exécution** de sauvegardes complètes.

Cumbersome structure. Some examples, although translated using grammatical structures, were classified as incorrect. These examples showed cumbersome structures that would not be natural in the target language. Please include in this group the instances you judge were classified as incorrect due to stylistic issues. See examples below:

EN: Profiles can be modified as required to meet the **changing** needs of user groups.
ES MT: Los perfiles se pueden modificar como sea necesario para cumplir las necesidades **que se modifican de** grupos de usuario.
Explanation: Despite the relative clause “que se modifican” being grammatical, it is not natural in Spanish because a more simple way, an adjective, exists to translated the -ing word. Basically, it is a similar difference between generating "changing needs of user groups" or "needs that are changing of user groups".
Correct version: Los perfiles se pueden modificar como sea necesario para cumplir las necesidades **cambiantes** de los grupos de usuarios.

Other. Please describe any other error affecting the –ing constituent for your target language in this column.

Comment. Please use the “comment” column if you want to make a comment on the errors marked in the list or to add any information you consider relevant for a complete understanding of the problems.

Layout of the excel file⁶¹:

Column A “source sentences” contains the English sentences that include -ing words which were classified as incorrect by evaluators. Note that the constituent of study is highlighted in red to help identify it easily.

Column B “raw MT output” contains the machine translated output of the source sentences. The source sentences were machine translated using the red highlighting and the MT preserved it where it thought it belonged in the target language. Please note that the highlight was kept only to help identify the constituents. It is not always accurate or even present in the target language examples.

Column C “post-edited version” contains a post-edited version of the raw MT output. This version has been provided as a possible final version. Evaluators could disagree with the PE version and consider the MT output correct or incorrect. This was provided to remind them that we were not aiming for a perfect output but for a version that is grammatical and accurate.

Use columns “T” to “W” to complete the analysis.

⁶¹ Please write your answers in the columns assigned for this purpose and do not modify the cells with content.

[illegible]

APPENDIX F: CLEANING GUIDELINES FOR THE FEATURE-BASED AUTOMATIC EVALUATION

The aim of the experiment we are carrying at the moment is to examine whether automatic metrics can be used to evaluate MT output for -ing words. Automatic metrics work by comparing the raw MT output against a reference translation - a post-edited version in our case - and calculate the similarity between them. The more similar they are, the better the quality of the raw MT output is considered to be. In order to test the usefulness of the automatic metrics, we are going to calculate the automatic scores for the 1,800 -ing words previously evaluated by human judges. Then we will be able to compare whether the results correlate or not.

Although automatic metrics are usually used for whole texts or sentences, we are going to test them at a subsentential level. This means that we have to input the exact strings we want the automatic metrics to compare. Because we are going to compare whether the judgement of the human evaluators matches with the results of the automatic metrics, we must ensure that the words considered by the human evaluators match with the string we will input for the automatic metrics. In order to do that, we have examined the evaluation guidelines and the Q&A list created during the human evaluation and we have a somewhat clear idea of which words they focused on and which ones they did not consider.

When we prepared the 1,800 examples for the human evaluation, we highlighted the -ing words in the source text in order to help evaluators identify them easily. Equally, we machine translated the source segments using the highlighting and therefore Systran included the highlight for the words it considered were the translation of the -ing word. What we are aiming to do is to use this highlighting to extract the exact strings from the raw MT output and the post-edited version as input for the automatic metrics. However, the highlighting was not always successfully competed by Systran.⁶² In many cases, the highlighting does not include all the words considered by humans. Therefore, what we are asking you to do is to follow the guidelines below to “clean” the highlighting of the raw MT output and the post-edited version.⁶³

⁶² Three different cases apply: (1) Systran did not highlight any word in the output; (2) Systran highlighted extra words in the output; (3) Systran did not highlight all the words that belong to the direct translation of the -ing word in the output.

⁶³ These guidelines were produced based on the evaluation guidelines and Q&A from the evaluation session, and the issues found in the cleaning of the English-Spanish and

The evaluators were asked to evaluate the translation of the -ing word highlighted in each example (grammaticality and accuracy). Therefore, and in order to decide what exactly to consider the translation of the -ing word:

Highlight all the words in the target language that contain the notions expressed within the -ing word such as meaning, tense, aspect, number/gender/etc. agreement, person (particularly for implicit subjects), etc.

Mainly bear in mind that:

We want to avoid the automatic metrics penalising a correct example just because the highlighting in the raw MT output and the post-edited version is different.

We want to make sure that the changes to convert the raw MT output into the post-edited version are not numbered up nor down because of inaccurate highlighting.

STRUCTURES	TARGET HIGHLIGHTING
direct translation structure	the complete structure to which the -ing word was translated should be highlighted. For instance, if an adjective is translated into a relative clause “that + verb”, “that” should also be translated
adverbials, catenatives, phrasal verbs, prepositional verbs	prepositions and subordinate conjunctions should not be highlighted
verbal regime	preposition given by the verbal phrase
adverbials – implicit subjects	the subject and verb should be highlighted for -ing words completing implicit subjects
progressive tenses	the whole verbal form should be highlighted
articles	missing/additional articles incorrect articles preceding the translation of an -ing word should be highlighted
compounds	when the translation of an -ing word forms a compound, the whole compound should be highlighted
attachments	should the translation of the -ing word have any pronoun attached, for instance, it should be highlighted
transpositions (change of part-of-speech from source to target)	include any where context is not transposed with them: translation of adverbs, determiners in front of nouns
anomalies	if a small anomaly such as a double prefix occurs, it should not be highlighted
position	if the translation of the -ing is split make sure there is a blank space between the different

English-French language pairs. The mismatches found in your target language might be different. Please, if in doubt, do not hesitate to contact me.

	parts
remember only 1 -ing word should be evaluated per example!	

Example:

<table><tr><td>source</td></tr><tr><td>Close cleaning media.</td></tr></table>		source	Close cleaning media.
source			
Close cleaning media.			
raw MT output	post-edited version		
Cierre los soportes de la limpieza .	Cierre los soportes de limpieza .		

Data as we have it at the moment. The -ing word to be evaluated is highlighted in the source. Systran has highlighted what it considered the translation of the -ing word into Spanish. The post-editor has highlighted what he/she considered the translation of the -ing word into Spanish.

raw MT output	post-edited version
Cierre los soportes de la limpieza .	Cierre los soportes de limpieza .

We would like you to modify the highlighting so that it matches with the words considered by the human evaluators following the guidelines above.

raw MT string	post-edited version string
de limpieza	de limpieza

We will then extract the highlighted words to obtain the strings to be compared by the automatic metrics.

As you can see in the example above, the highlighted -ing word is a participial adjective which is modifying the noun *media*. This participial adjective has been successfully translated as a post-modifying constituent into Spanish *de la limpieza*. However, Systran did not highlight the whole post-modifier, in contrast with the post-editor. Also, the post-editor removed the article *la*, which, although grammatical, it would not be natural in Spanish. Because we asked evaluators to judge whether an -ing word was grammatical and accurate without considering stylistic preferences, the article should not be highlighted. If the article remains highlighted in the raw MT

output and it does not appear in the post-edited version, automatic metrics will consider that the strings are not identical and will penalise the machine translation.

APPENDIX G: HUMAN EVALUATION AND AUTOMATIC METRICS CORRELATIONS

SPANISH - Correlations

		HUMAN	TER	EditD	GTM	METEOR	NIST
HUMAN	Pearson Correlation	1	-.691(**)	-.574(**)	.642(**)	.561(**)	.510(**)
	Spearman's rho Correlation Coefficient	1.000	-.681(**)	-.630(**)	.602(**)	.541(**)	.386(**)
	Kendall's tau_b Correlation Coefficient	1.000	-.626(**)	-.555(**)	.555(**)	.490(**)	.293(**)
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	1800	1800	1800	1800	1800	1800
TER	Pearson Correlation	-.691(**)	1	.760(**)	-.798(**)	-.804(**)	-.724(**)
	Spearman's rho Correlation Coefficient	-.681(**)	1.000	.929(**)	-.883(**)	-.839(**)	-.720(**)
	Kendall's tau_b Correlation Coefficient	-.626(**)	1.000	.867(**)	-.860(**)	-.803(**)	-.598(**)
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	1800	1800	1800	1800	1800	1800
EditD	Pearson Correlation	-.574(**)	.760(**)	1	-.653(**)	-.741(**)	-.612(**)
	Spearman's rho Correlation Coefficient	-.630(**)	.929(**)	1.000	-.971(**)	-.814(**)	-.687(**)
	Kendall's tau_b Correlation Coefficient	-.555(**)	.867(**)	1.000	-.887(**)	-.749(**)	-.546(**)
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	1800	1800	1800	1800	1800	1800
GTM	Pearson Correlation	.642(**)	-.798(**)	-.653(**)	1	.679(**)	.720(**)
	Spearman's rho Correlation Coefficient	.602(**)	-.883(**)	-.971(**)	1.000	.724(**)	.671(**)
	Kendall's tau_b Correlation Coefficient	.555(**)	-.860(**)	-.887(**)	1.000	.700(**)	.567(**)
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	1800	1800	1800	1800	1800	1800
METEOR	Pearson Correlation	.561(**)	-.804(**)	-.741(**)	.679(**)	1	.638(**)
	Spearman's rho Correlation Coefficient	.541(**)	-.839(**)	-.814(**)	.724(**)	1.000	.590(**)
	Kendall's tau_b Correlation Coefficient	.490(**)	-.803(**)	-.749(**)	.700(**)	1.000	.488(**)
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	1800	1800	1800	1800	1800	1800
NIST	Pearson Correlation	.510(**)	-.724(**)	-.612(**)	.720(**)	.638(**)	1
	Spearman's rho Correlation	.386(**)	-.720(**)	-.687(**)	.671(**)	.590(**)	1.000

	Coefficient						
	Kendall's tau_b Correlation Coefficient	.293(**)	-.598(**)	-.546(**)	.567(**)	.488(**)	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	1800	1800	1800	1800	1800	1800

** Correlation is significant at the 0.01 level (2-tailed).

FRENCH - Correlations

		HUMAN	TER	EditD	GTM	METEOR	NIST
HUMAN	Pearson Correlation	1	-.595(**)	-.519(**)	.603(**)	.529(**)	.564(**)
	Spearman's rho Correlation Coefficient	1.000	-.623(**)	-.579(**)	.597(**)	.549(**)	.571(**)
	Kendall's tau_b Correlation Coefficient	1.000	-.546(**)	-.469(**)	.534(**)	.480(**)	.475(**)
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	1800	1800	1800	1800	1800	1800
TER	Pearson Correlation	-.595(**)	1	.687(**)	-.823(**)	-.666(**)	-.764(**)
	Spearman's rho Correlation Coefficient	-.623(**)	1.000	.817(**)	-.927(**)	-.708(**)	-.905(**)
	Kendall's tau_b Correlation Coefficient	-.546(**)	1.000	.697(**)	-.888(**)	-.657(**)	-.806(**)
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	1800	1800	1800	1800	1800	1800
EditD	Pearson Correlation	-.519(**)	.687(**)	1	-.666(**)	-.763(**)	-.647(**)
	Spearman's rho Correlation Coefficient	-.579(**)	.817(**)	1.000	-.798(**)	-.847(**)	-.773(**)
	Kendall's tau_b Correlation Coefficient	-.469(**)	.697(**)	1.000	-.675(**)	-.729(**)	-.608(**)
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	1800	1800	1800	1800	1800	1800
GTM	Pearson Correlation	.603(**)	-.823(**)	-.666(**)	1	.664(**)	.920(**)
	Spearman's rho Correlation Coefficient	.597(**)	-.927(**)	-.798(**)	1.000	.658(**)	.966(**)
	Kendall's tau_b Correlation Coefficient	.534(**)	-.888(**)	-.675(**)	1.000	.632(**)	.895(**)
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	1800	1800	1800	1800	1800	1800
METEOR	Pearson Correlation	.529(**)	-.666(**)	-.763(**)	.664(**)	1	.644(**)
	Spearman's rho Correlation Coefficient	.549(**)	-.708(**)	-.847(**)	.658(**)	1.000	.640(**)

	Kendall's tau_b Correlation Coefficient	.480(**)	-.657(**)	-.729(**)	.632(**)	1.000	.574(**)
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	1800	1800	1800	1800	1800	1800
	Pearson Correlation	.564(**)	-.764(**)	-.647(**)	.920(**)	.644(**)	1
	Spearman's rho Correlation Coefficient	.571(**)	-.905(**)	-.773(**)	.966(**)	.640(**)	1.000
NIST	Kendall's tau_b Correlation Coefficient	.475(**)	-.806(**)	-.608(**)	.895(**)	.574(**)	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	1800	1800	1800	1800	1800	1800
	Pearson Correlation	.564(**)	-.764(**)	-.647(**)	.920(**)	.644(**)	1
	Spearman's rho Correlation Coefficient	.571(**)	-.905(**)	-.773(**)	.966(**)	.640(**)	1.000

** Correlation is significant at the 0.01 level (2-tailed).

GERMAN - Correlations

		HUMAN	TER	EditD	GTM	METEOR	NIST
HUMAN	Pearson Correlation	1	-.484(**)	-.527(**)	.528(**)	.469(**)	.424(**)
	Spearman's rho Correlation Coefficient	1.000	-.544(**)	-.595(**)	.554(**)	.482(**)	.389(**)
	Kendall's tau_b Correlation Coefficient	1.000	-.502(**)	-.523(**)	.515(**)	.425(**)	.313(**)
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	1798	1798	1798	1798	1798	1798
TER	Pearson Correlation	-.484(**)	1	.597(**)	-.926(**)	-.796(**)	-.748(**)
	Spearman's rho Correlation Coefficient	-.544(**)	1.000	.943(**)	-.979(**)	-.817(**)	-.758(**)
	Kendall's tau_b Correlation Coefficient	-.502(**)	1.000	.840(**)	-.962(**)	-.753(**)	-.643(**)
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	1798	1800	1800	1800	1800	1800
EditD	Pearson Correlation	-.527(**)	.597(**)	1	-.625(**)	-.686(**)	-.579(**)
	Spearman's rho Correlation Coefficient	-.595(**)	.943(**)	1.000	-.945(**)	-.872(**)	-.766(**)
	Kendall's tau_b Correlation Coefficient	-.523(**)	.840(**)	1.000	-.854(**)	-.761(**)	-.603(**)
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	1798	1800	1800	1800	1800	1800
GTM	Pearson Correlation	.528(**)	-.926(**)	-.625(**)	1	.819(**)	.793(**)
	Spearman's rho Correlation Coefficient	.554(**)	-.979(**)	-.945(**)	1.000	.820(**)	.771(**)
	Kendall's tau_b Correlation	.515(**)	-.962(**)	-.854(**)	1.000	.769(**)	.663(**)

METEOR	Coefficient						
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	1798	1800	1800	1800	1800	1800
	Pearson Correlation	.469(**)	-.796(**)	-.686(**)	.819(**)	1	.783(**)
	Spearman's rho Correlation Coefficient	.482(**)	-.817(**)	-.872(**)	.820(**)	1.000	.757(**)
	Kendall's tau_b Correlation Coefficient	.425(**)	-.753(**)	-.761(**)	.769(**)	1.000	.615(**)
	Sig. (2-tailed)	.000	.000	.000	.000		.000
NIST	N	1798	1800	1800	1800	1800	1800
	Pearson Correlation	.424(**)	-.748(**)	-.579(**)	.793(**)	.783(**)	1
	Spearman's rho Correlation Coefficient	.389(**)	-.758(**)	-.766(**)	.771(**)	.757(**)	1.000
	Kendall's tau_b Correlation Coefficient	.313(**)	-.643(**)	-.603(**)	.663(**)	.615(**)	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	1798	1800	1800	1800	1800	1800

** Correlation is significant at the 0.01 level (2-tailed).

JAPANESE - Correlations

		HUMAN	TER	EditD	GTM	NIST
HUMAN	Pearson Correlation	1	-.533(**)	-.549(**)	.613(**)	.496(**)
	Spearman's rho Correlation Coefficient	1.000	-.654(**)	-.646(**)	.661(**)	.482(**)
	Kendall's tau_b Correlation Coefficient	1.000	-.577(**)	-.568(**)	.589(**)	.375(**)
	Sig. (2-tailed)		.000	.000	.000	.000
	N	1800	1800	1800	1800	1800
TER	Pearson Correlation	-.533(**)	1	.757(**)	-.833(**)	-.654(**)
	Spearman's rho Correlation Coefficient	-.654(**)	1.000	.966(**)	-.980(**)	-.755(**)
	Kendall's tau_b Correlation Coefficient	-.577(**)	1.000	.888(**)	-.919(**)	-.605(**)
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1800	1800	1800	1800	1800
EditD	Pearson Correlation	-.549(**)	.757(**)	1	-.846(**)	-.685(**)
	Spearman's rho Correlation Coefficient	-.646(**)	.966(**)	1.000	-.970(**)	-.749(**)
	Kendall's tau_b Correlation Coefficient	-.568(**)	.888(**)	1.000	-.898(**)	-.598(**)
	Sig. (2-tailed)	.000	.000		.000	.000

	N	1800	1800	1800	1800	1800
GTM	Pearson Correlation	.613(**)	-.833(**)	-.846(**)	1	.772(**)
	Spearman's rho					
	Correlation	.661(**)	-.980(**)	-.970(**)	1.000	.777(**)
	Coefficient					
	Kendall's tau_b					
	Correlation	.589(**)	-.919(**)	-.898(**)	1.000	.644(**)
	Coefficient					
	Sig. (2-tailed)	.000	.000	.000		.000
	N	1800	1800	1800	1800	1800
NIST	Pearson Correlation	.496(**)	-.654(**)	-.685(**)	.772(**)	1
	Spearman's rho					
	Correlation	.482(**)	-.755(**)	-.749(**)	.777(**)	1.000
	Coefficient					
	Kendall's tau_b					
	Correlation	.375(**)	-.605(**)	-.598(**)	.644(**)	1.000
	Coefficient					
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	1800	1800	1800	1800	1800

** Correlation is significant at the 0.01 level (2-tailed).

APPENDIX H: CL TRANSLATION EFFECT EVALUATION GUIDELINES

In the Evaluation Pack you will find two tasks to complete. First of all, we will ask you to fill in a brief questionnaire (questionnaire.doc). Secondly, you will find the files required for an evaluation. The evaluation involves the judging of the grammaticality and accuracy of certain structures. You will find more specific guidelines for the evaluation task below.

Please read through all the instructions before commencing.

The aim of this evaluation is to analyse the machine translation of specific grammatical structures to gather information about the efficiency with which our rule-based machine translation (MT) system handles them.

An Excel file (evaluation_XX.xls) is provided with a layout for evaluation. The file contains 163 segments. In each segment, only one structure has been highlighted for you to evaluate. You must not evaluate the whole segment.

1. Read the source segment in Column A and identify the highlighted structure.
2. Read the machine translation output in Column B and identify the translation of the structure highlighted in the source.

The MT system has transferred the source segment highlighting to the target segment. However, this process is not always precise. Still, the highlighting was kept to help you identify the translations easily. Please use the highlighting as a guide only.

3. Evaluate whether the machine translation of the structure highlighted in the source is **grammatical** and **accurate**.

By grammatical we mean that the output respects the reference grammatical rules of the target language. By accurate we mean that the information contained in the source structure has been reproduced without distortion in the translation.

If both requirements are true for your target language, the segment should be considered "Correct" (Column C). If any of the requirements is not true, the segment should be considered "Incorrect" (Column D). Please note that only the translation corresponding to the highlighted source structure should be judged. Type an "X" in the appropriate cell in the Excel file provided. Please use only the fields assigned for evaluation and do not modify or change the order of the cells with content.

Remember that we are not aiming for a perfect output. Terminological precision should be overlooked. We are looking for a translation that is grammatical and accurate.

Microsoft Excel - evaluation_ES.xls				
Type a question for help				
C2 A B C D				
1	SOURCE SENTENCES	SPANISH - MACHINE TRANSLATION OUTPUT	CORRECT	INCORRECT
2	If Unix files are included in the selections, the characters /u should appear before the comma that separates the entries, and the last quotation mark.	Si los ficheros de Unix se incluyen en las selecciones, los caracteres /u deben aparecer antes de la coma que separa las entradas, y de la marca de cita pasada.		
3	When you redirect Exchange data, the service pack on the Exchange server where data is being redirected should be the same as the service pack on the original Exchange server.	Cuando usted vuelve a dirigir los datos de Exchange, el paquete de servicios en el servidor de Exchange donde se están volviendo a dirigir los datos debe ser igual que el paquete de servicios en el servidor original de Exchange.		
4	The product eliminates cumbersome and time-consuming tasks that face database administrators, thereby reducing costs.	El producto elimina las tareas incómodas y desperdiciadoras de tiempo que hacen frente a administradores de base de datos, de tal modo reduciendo costes.		
5	The specified disk storage limit was reached when you attempted to add a new revision to the desktop user data folder.	El límite especificado del almacenamiento en discos fue alcanzado cuando usted intentó agregar una nueva revisión a la carpeta de escritorio de los datos de utilizador.		
6	Create a cleaning job	Cree un trabajo de la limpieza		
7	Type the name of a computer that is running the Backup Exec Remote Agent for Windows Systems, or click Browse to navigate to the remote agent computer.	Pulse el nombre de un ordenador que esté funcionando con el Backup Exec Remote Agent para Windows Systems, o haga clic Browse para navegar al ordenador alejado del agente.		
8	After you install Backup Exec on the media server, Backup Exec Workstation Agent software can be installed and configured on remote workstations on the network.	Después de que usted instale Backup Exec en los media servidor, el software de Backup Exec Workstation Agent se puede instalar y configurar en sitios de trabajo alejados en la red.		
9	Specifies the amount of time Backup Exec should wait for the job to complete before Backup Exec cancels the job.	Especifica la cantidad de tiempo Backup Exec debe esperar el trabajo de terminar antes de que Backup Exec cancele el trabajo.		
	The product eliminates cumbersome and time-consuming tasks that face database administrators, thereby reducing costs.	El producto elimina las tareas incómodas y desperdiciadoras de tiempo que hacen frente a administradores de base de datos, de tal modo reduciendo		

Before submitting, please check that an answer has been provided for all the segments.

Thank you!

APPENDIX I: AUTOMATIC SOURCE RE-WRITING, GLOBAL S&R, & SPE TRANSLATION EFFECT EVALUATION GUIDELINES

For this evaluation you will be asked to perform two separate tasks regarding the machine translation of -ing words into your mother tongue.

Please read through all the instructions first before commencing the tasks.

Task 1 – -ing groups

1. Open the Excel file called “ING_evaluation_XX” and select the “ING_groups” worksheet.
2. There are a number of sentences for which the translation of the -ing word must be evaluated.
3. Read the Source sentence and identify the relevant -ing word. This is highlighted to help you in the task.
4. Read MT1 and MT2 and identify the relevant part that corresponds to the -ing translation. In order to help you in this task, the translation of the -ing word is identified in MT1. Please note that this highlighting may not be accurate or even present as it was placed automatically by the MT system.
5. Compare the -ing translations for MT1 and MT2 and evaluate which is better in terms of grammaticality and accuracy. By grammatical we mean that the output respects the reference grammatical rules of the target language. By accurate we mean that the information contained in the source structure has been reproduced without distortion in the translation. Please make your judgement based on the translation of the -ing word only, not the whole sentence. You are given the following options:
 - a. MT1 is better. Select MT1 from the drop-down menu.
 - b. MT2 is better. Select MT2 from the drop-down menu.
 - c. None is better than the other. Select SAME from the drop-down menu.
6. The sentences are grouped according to the type of -ing word to be evaluated. Please rate each group in one session.
7. If you would like to make any comments about your judgements, you can do so in Column E “Comments”, but this is not compulsory.
8. Please save your results. Before submitting, please check that an answer has been provided for all the segments.

Microsoft Excel - ING_evaluation_FR.xls				
File Edit View Insert Format Tools Data Window Help				
Type a question for help				
D2				
	A	B	C	D
	SOURCE SENTENCE	MT1	MT2	Best ING translation
1				
2	Creating a policy	Création une politique	Création d'une politique	
3	Starting Backup Exec after installing the Library Expansion Option	Démarrant Backup Exec après avoir installé Library Expansion Option	Démarrer de Backup Exec après avoir installé Library Expansion Option	MT1 MT2 SAME
4	Pending Jobs	En attendant des travaux	Installation du contrôle de la BV pour Microsoft SQL Server	
5	Installing bv Control for Microsoft SQL Server	En installant la BV contrôlez pour Microsoft SQL Server	Suppression des règles de modèle	
6	Renaming robotic libraries and drives	Renommer des bandothèques et des lecteurs	Renommer des bandothèques et des lecteurs	
7	Using Windows' Automated System Recovery and System Restore to recover a Windows XP or Windows Server 2003 system	Utiliser la récupération automatique du système et la restauration du système de Windows pour récupérer un système de Windows XP ou de Windows Server 2003	L'utilisez de la récupération automatique du système et de la restauration du système de Windows pour récupérer un système de Windows XP ou de Windows Server 2003	
8	Restoring Exchange data from snapshot backups	Données Exchange de Restauration des sauvegardes de clichés	Restauration des données Exchange des sauvegardes de clichés	

Task 2 – random -ing words

1. Open the Excel file called “ING_evaluation_XX” and select the “ING_random” worksheet.
2. There are a number of sentences to be evaluated according to two different questions.
3. Question 1: Which translation is better in terms of grammaticality and accuracy?
 - a. Read the Source sentence.
 - b. Read MT1 and MT2.
 - c. Compare MT1 and MT2 and evaluate which is better in terms of grammaticality and accuracy. By grammatical we mean that the output respects the reference grammatical rules of the target language. By accurate we mean that the information contained in the source structure has been reproduced without distortion in the translation. Please make your judgement based on the quality of the whole sentence. You are given the following options:
 - i. MT1 is better. Select MT1 from the drop-down menu.
 - ii. MT2 is better. Select MT2 from the drop-down menu.
 - iii. None is better than the other. Select SAME from the drop-down menu.
4. Question 2: Which -ing translation is better in terms of grammaticality and accuracy?
 - a. Read the Source sentence and identify the relevant -ing word. This is highlighted to help you in the task.
 - b. Read MT1 and MT2 and identify the relevant part that corresponds to the -ing translation. In order to help you in this task, MT1 identified the translation of the -ing word. Please note that this highlighting may not be accurate or even present as it was placed automatically by the MT system.
 - c. Compare the -ing translations for MT1 and MT2 and evaluate which is better in terms of grammaticality and accuracy. By grammatical we mean that the output respects the reference grammatical rules of the target language. By accurate we mean that the information contained in the source structure has been reproduced without distortion in the translation. Please make your judgement based on the translation of the -ing word only, not the whole sentence. You are given the following options:
 - iv. MT1 is better. Select MT1 from the drop-down menu.
 - v. MT2 is better. Select MT2 from the drop-down menu.
 - vi. None is better than the other. Select SAME from the drop-down menu.
5. If you would like to make any comments about your judgements, you can do so in Column F “Comments”, but this is not compulsory.
6. Please save your results. Before submitting, please check that an answer has been provided for all the segments.

Microsoft Excel - ING_evaluation_FR.xls					
Type a question for help					
11 113% Arial					
TV Reply with Changes... Exp Review...					
	A	B	C	D	E
	SOURCE SENTENCE	MT1	MT2	Best sentence	Best ING translation
	Comments				
1	For example, after installing the operating system on a crashed computer, you could restore a previous full backup of the system without worrying about overwriting later versions of operating system files.	Par exemple, après avoir installé le système d'exploitation sur un ordinateur tombé en panne, vous pourriez restaurer une sauvegarde complète précédente du système sans inquiéter les versions postérieures les écrasant à propos de des fichiers du système d'exploitation.	Par exemple, après avoir installé le système d'exploitation sur un ordinateur tombé en panne, vous pouvez restaurer une sauvegarde complète précédente du système sans préoccuper les versions plateforme les écrasant à propos des fichiers du système d'exploitation.		
2					
3	"Deleting a media vault" on page 301	« Supprimant un centre de sauvegarde » à la page 301	"Supprimant un centre de sauvegarde" MT1 MT2 SAME		
4	Understanding backup methods and their advantages	Méthodes de sauvegarde de compréhension et leurs avantages	Méthodes de sauvegarde de connaissance et leurs avantages		
5	Backup Exec supports IDR of Citrix Metaframe 1.8, XPa, XPe, and XP computers with the following exceptions:	Backup Exec prend en charge l'IDR de Citrix Metaframe 1.8, de XPa, de XPe, et d'ordinateurs de XP à les exceptions suivantes :	Backup Exec prend en charge l'IDR 1.8 Metaframe 1.8, XPa, XPe et d'ordinateurs XP à les exceptions suivantes :		
6	In a default installation, Backup Exec log files reside in the Logs directory, in the following path.	À une installation par défaut, les fichiers de consignation de journal Backup Exec résident dans le répertoire Journaux, dans le chemin d'accès suivant :	À une installation par défaut, les fichiers de consignation de journal Backup Exec résident en général dans le répertoire Journaux, dans le chemin d'accès suivant :		
	The SPSI Information tab of a website library.	Il fournit l'information SPSI de la bibliothèque.	Il fournit l'information SPSI de la bibliothèque.		

Thank you very much!

