Evaluation of Machine Translation Systems: The Translation Quality of Three Arabic Systems

A thesis
Submitted to the Council of the College of Arts of Al-Mustansiriyah University
in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Linguistics and Translation

By
Yasmin Hikmat Abdul-Hamid Hannouna

Supervised by
Assist. Prof. Dr. Aliya M. H. Ahmed Al-Rubaisi
Assist. Prof. Dr. Mu'ayad A. R. Firdhil Al-Ubaydi

March 2004

Muharam 1425
Evaluation of Machine Translation Systems: The Translation Quality of Three Arabic Systems

A THESIS
SUBMITTED TO THE COUNCIL OF THE
COLLEGE OF ARTS OF AL-MUSTANSIRIYYA UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF PHILOSOPHY IN LINGUISTICS AND TRANSLATION

By

YASMIN HIKMET ABDUL-HAMID HANNOUNA

Supervised by

Asst. Prof. Dr. Alya’ M. H. Ahmad Al-Rubai’i
Asst. Prof. Dr. Mu’ayad A. R. Fadhil Al-’Ubaydi

March 2004

Muharam 1425
We certify that this thesis was prepared under our supervision at Al-Mustansiriyyah University in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Linguistics and Translation.

Signature

Supervisor: Asst. Prof. Dr. Alya’ M. H. Ahmad Al-Rubai’i
Date:

Signature

Supervisor: Asst. Prof. Dr. Mua’yad A. R. Fadhil Al-’Ubaydi
Date:

In view of the available recommendation, I forward this thesis for debate by the Examining Committee.

Signature

Name: Dr. Sedeer K. M. Al-Hadithi

Head of the Dept. of Translation
College of Arts, Al-Mustansiriyyah University
Date:

We certify that we have read this thesis and as Examining Committee examined the student in its content and that, in our opinion, it is adequate with Excellent standing as a thesis for the degree of Doctor of Philosophy in Linguistics and Translation.

Signature

Name:
Date:
Member                                      Member

Signature                                   Signature
Name:                                       Name:
Date :                                      Date :

Member                                      Member

Signature                                   Signature
Name:                                       Name:
Date :                                      Date :

Chairman
Approved by the Council of the College of Arts

Signature
Name: Prof. Dr. Hassan Al-Khafaji
Dean of the College of Arts,
Al-Mustansiriyya University
Date:
“MT Evaluation is pain, expensive, stressful and inconclusive. What’s needed: a project of evaluating evaluations.”

Eduard Hovy (1999)
The British Government is said to be looking at the possibilities of a British machine translation system (February 1986 *Language Monthly*)
DEDICATION

TO MY FAMILY WITH LOVE
AND
TO ALL WHO SEEK
TO IMPROVE ARABIC MT
WITH MUCH RESPECT
ACKNOWLEDGEMENTS

Acknowledgments can never be made to all those who have nourished one’s intellectual life. Chief among these, in my case, are my supervisors Asst. Prof. Dr. Alya’ M. H. Ahmed Al-Rubai’i and Asst. Prof. Dr. Mu’ayad A. R. Fadhil Al-‘Ubaydi who have given me sustenance and have kindled the desire for more knowledge. I am deeply indebted to them for everything they did, so that the present work can see the light.

I would like to express my gratitude to the following persons for their help, generous hospitality and provision of references: Dr. Muna Abdul-Sahib Al-Alwan, Dr. Muhammed Kadoomi and Dr. Muhammed Riyadh Al-Jazzar/ Dept. of European Languages/College of Arts/ University of Bahrain; Prof. Abdullah Al-Shunnaq and Dr. Muhammed Al-Sarairah/ Dept. of Translation / College of Arts / Yarmook University.

A particular debt of thanks and appreciation goes to Dr. Ahmed Guessoum/ Dept. of Computer Science/ University of Sharjah; Prof. Eduard Hovy/ Information Science Institute/ University of Southern California and Prof. Andrei Popescu-Belis/ ISSCO/ TIM/ETI / University of Geneva for their constant encouragement and invaluable comments and suggestions on my work through e-mails, and for sending me many valuable references which have greatly contributed to the progress of this study.

I am also deeply appreciative and sincerely grateful to Prof. Mona Baker/UMIST/University of Manchester for her continuous encouragement and support and for sending me very useful references.

Special words of thanks and appreciation are due to Ms. Sandra Menzi/ISSCO/University of Geneva and Prof. Harold Somers/UMIST/ University of Manchester for sending me important references; Prof. Abdul-Wahid Lu’lu’a/ Dept. of English /Philadelphia University for his help and constant encouragement. Special thanks are also due to the librarians at the Jordanian University, Ms.Afaaf/ Chief Librarian / Batraa’ University, and Mr.Awadh and Ms.Baryooni / Central Library / Yarmook University for their kindness and generous help.

I would like to thank the members of the jury (See Appendix B) who have so carefully considered the model of the present evaluation, the criteria selected and the scoring scheme of finding out the total system’s performance, and made valuable suggestions and comments.
My sincere respect and gratitude also goes to Prof. Gaasid Yasir Al-Zaydi/Dept. of Arabic/College of Education for Women/Baghdad University for proof reading the Arabic reference translation of the sample texts in this study.

I am also particularly indebted to Dr. Baydaa’ Al-‘Ubaydi/Dept of Translation/ College of Arts/University of Mustansiriyya for her support and generous help in acting as a second evaluator for the tests of the experimental work.

I would like to express my gratitude to Mr. Hasan Kasim/Dept. of Computer Science/College of Science for Women/University of Baghdad for his generous help and support and provision of useful references.

Finally, in stressful moments of struggle in the writing of this work, I shall never forget the inspirational help, patience and forbearance of my father and my lovely sister Menahil who really deserve my warm thanks and respect. God bless them.
ABSTRACT

Evaluation is an implicit aspect of all human activity. With respect to MT, it remains an open fundamental issue and one of the most important stages in the life cycle of an MT system.

The present evaluation study investigates the overall quality of three currently available English-into-Arabic MT systems. The evaluation deals with selected quality characteristics and various text types. This is to bear on their ability and extent to satisfy specific requirements and help users to fulfil their tasks. It is contended at the outset of the study that the inadequacy of the lexicon and the linguistic components of an MT system affects the quality of the whole system. As such, the evaluation of the cognitive and linguistic capabilities of an MT system is clearly a pivotal issue for the process of evaluating MT systems.

Two interrelated aspects of MTE have been identified and discussed: a.) the framework model of MTE and b.) the important stages of the evaluation process. The quality requirements (both functional and non-functional) have been expressed in terms of quality characteristics and sub-characteristics. Much attention has been paid to the functional criteria in this study as they are the most common and essential in every evaluation process to determine the quality of an MT output.

The theoretical construct adopted in this study is based on the Framework for Evaluation of MT in ISLE (FEMTI, 2003) which is the most recent and comprehensive model of evaluation. The study focuses on the selected functional and computational criteria of system’s performance. For each criterion under evaluation, the appropriate metrics, scoring scheme and methods regarding the overall judgment of the product have been identified and described.

The evaluation, therefore, constitutes a standard test bed application of this methodology (i.e., task-oriented testing and benchmark testing). The proposed model for the functional criteria is a black-box type, comparative and adequacy-oriented evaluation. As for the non-functional criteria, the evaluation model is said to be the comparative performance and adequacy-oriented type. The sample represents a total of 268 English sentences taken from twelve various special-domain texts. They have been used as input testing data in the evaluation of selected functional criteria of the three MT systems. Some computational criteria have also been evaluated. These systems have been tested under experimental conditions by two evaluators. The end-user and his tasks have been identified and discussed in relation to the ISLE characteristics of the translation quality.
Detailed analyses and classification of the results concerning the selected criteria are presented with Excel tables, charts and graphs. The overall comparison of the three systems in terms of quality assessment of both criteria and texts level confirm the following generalizations:

1.) English-into-Arabic MT systems suffer from serious drawbacks especially related to the grammar and meanings of the translated sentence. Their output reflects many deficiencies in translating various text types and they all need serious improvements.

2.) All the three MT systems considered produce average or below average quality, i.e., with a total quality performance of 58% for MA, 45% for NA and 44% for GW.

3.) MA system is better than the others in performing scientific translations and the three systems only manage commercial and computational texts with a satisfactory translation.

4.) The end user can use these systems to grasp the general idea of the ST, or translate short and simple texts.

5.) Three major types of problems have been identified with these systems: a.) cognitive, b.) linguistic and c.) operational. The results of these problems indicate that:

i.) The Arabic MT output is, in general, unclear and unintelligible. It also suffers from being unfaithful in conveyance of SL information.

ii.) The major linguistic problems in the data centre around inadequate terminology interpretation, incorrect rendition of SL grammatical structure and erroneously recognized TL morphology.

iii.) The operational problems are attributed to certain impediments in measuring the speed of translation and some limitations relevant to the design and performance of the user dictionaries of these systems.

Having identified these problems, the researcher then investigates their possible sources. Some sources are related to variation between SL and TL linguistic structures and deficiencies of system’s lexicon and components. Others concern the lack of awareness of the strategic nature of Arabic MT, and the significance of MTE.

On the basis of these findings, a number of suggestions and recommendations are made. These can be of assistance to systems’ designers and developers to find ways of improvement. Further, some suggestions are useful to the potential users of these systems to determine which of these systems can best satisfy their needs.
TABLE OF CONTENTS

Subject ................................................. Page
DEDICATION ................................................ vi
ACKNOWLEDGEMENTS ..................................... vii
ABSTRACT .................................................. ix
TABLE OF CONTENTS ......................................... xii
KEY OF ABBREVIATIONS AND ACRONYMS ................... xviii
UNKNOWN ABBREVIATIONS AND ACRONYMS ................. xx
LIST OF TABLES ............................................ xxi
LIST OF FIGURES ........................................... xxiii

CHAPTER ONE: INTRODUCTION
1.1 Preliminaries ............................................. 1
  1.1.1 Evaluation of Machine Translation Systems .......... 1
  1.1.2 Evaluation of Existing Arabic Systems ............... 4
1.2 Statement of the Problem .................................. 4
1.3 Hypotheses .............................................. 5
1.4 Aims ..................................................... 7
1.5 Scope .................................................... 7
1.6 Procedure ............................................... 8
1.7 Significance ............................................. 9
1.8 Plan ...................................................... 11

CHAPTER TWO: EVALUATING MT SYSTEMS: SOME
BASIC ISSUES
2.1 General Overview ........................................ 13
2.2 Scope of Evaluation ................................................................. 15
2.3 The Role of Evaluation in MT ................................................... 17
2.4 Some History ........................................................................... 18
  2.4.1 Background ........................................................................ 18
  2.4.2 Some Past Evaluations .......................................................... 20
    2.4.2.1 ALPAC ......................................................................... 20
    2.4.2.2 (D) ARPA ................................................................. 21
    2.4.2.3 The Hewlett Packard Laboratories Study ....................... 23
    2.4.2.4 MUC-3 ....................................................................... 23
    2.4.2.5 COBALT ..................................................................... 25
2.5 Standards Work Related to Evaluation ........................................ 25
  2.5.1 Previous Approaches to MTE ............................................... 27
    2.5.1.1 JEIDA (1992) .............................................................. 27
    2.5.1.2 The EAGLES Evaluation Guidelines ............................. 28
    2.5.1.3 The EAGLES 7-Step Recipe ......................................... 30
    2.5.1.4 The ISO/IEC Standards for Software Evaluation .......... 32
      2.5.1.4.1 A Growing Set of Standards ................................. 32
      2.5.1.4.2 The Definition of a Quality Model ......................... 33
      2.5.1.4.3 Stages in the Evaluation Process ........................... 34
    2.5.1.5 The ISLE MTE Framework ......................................... 35
      2.5.1.5.1 Origins and First Version .................................... 36
      2.5.1.5.2 Key Principle of the Second Draft ....................... 37
      2.5.1.5.3 A Framework for the Evaluation of Machine
                   Translation in ISLE (2003) ................................. 37
      2.5.1.5.4 Analyzing the Behaviour of Measures ................... 39
2.6 New Trends and Future Work in MTE ...................................... 41
2.7 Formal and Functional Approaches to MTE ............................... 42
2.8 Evaluation for MT Stakeholders ......................................................... 44
2.9 Types and Purposes of MTE ................................................................. 46
  2.9.1 Feasibility Evaluation ................................................................. 46
  2.9.2 Internal Evaluation ................................................................... 47
  2.9.3 Black-box Vs. Glass-box Evaluation .......................................... 48
  2.9.4 Diagnostic Evaluation ................................................................. 48
  2.9.5 Declarative Evaluation ................................................................. 48
  2.9.6 Adequacy Evaluation .................................................................. 49
  2.9.7 Performance Evaluation .............................................................. 49
  2.9.8 Operational Evaluation ............................................................... 50
  2.9.9 Usability Evaluation .................................................................... 50
  2.9.10 Comparison Evaluation ............................................................. 51
2.10 The Object of Evaluation ................................................................. 51
2.11 MT Systems ..................................................................................... 53
  2.11.1 Background ................................................................................ 53
  2.11.2 MT Systems and Text Type Considerations ................................ 54
  2.11.3 Classification of MT Systems ...................................................... 55
    2.11.3.1 Typology of MT Systems by Text Type ................................ 55
    2.11.3.2 Typology of MT Systems by Degree of Automation .......... 56
  2.11.4 Description of Arabic MT Systems ............................................. 57
    2.11.4.1 Sakhar .................................................................................. 58
    2.11.4.2 Turjman ............................................................................... 58
    2.11.4.3 ArabTrans ............................................................................ 58
    2.11.4.4 AL-Wafi ............................................................................... 60
    2.11.4.5 SYSTRAN ............................................................................ 61
    2.11.4.6 Applied Technology (Apptek) .............................................. 62
    2.11.4.7 Other Arabic MT Systems and Projects ................................ 62
2.11.4.7.1 Electronic Research Institute: The National Research Council in Cairo .................. 63
2.11.4.7.2 The Scientific Center in Cairo (IBM)......... 63
2.11.4.7.3 XEROX Company ................................. 63
2.11.4.7.4 Luts Company.................................. 63
2.11.4.7.5 The House of Computerizing the Arabic Text ......................................................... 64
2.11.4.7.6 Microsoft............................................. 64

2.11.4.8 Arabic MT Systems under Evaluation.............. 64
2.11.4.8.1 AL-Mutarjim AL-Arabey ............................ 64
  2.11.4.8.1.1 General ......................................... 64
  2.11.4.8.1.2 The General Dictionary ................. 67
  2.11.4.8.1.3 How Does AL-Mutarjim AL-Arabey Work ? .................. 67
  2.11.4.8.1.4 General Issues............................... 69
  2.11.4.8.1.5 AL-Mutarjim AL-Arabey 3.00 (2002) ................. 71
  2.11.4.8.2 An-Nakel AL-Arabi................................ 75
    2.11.4.8.2.1 General ...................................... 75
    2.11.4.8.2.2 How It Works ? ....................... 77
    2.11.4.8.2.3 Levels of System Operation ... 78
  2.11.4.8.3 Golden Al-Wafi 1.00 (2002) .................... 86

2.12 The State of Arabic Machine Translation ...................... 90
  2.12.1 Existing Arabic MT Systems and Arabic MTE .............. 91

2.13 Arabic MT and the Internet ........................................... 92

2.14 Recent Research on MTE ............................................ 95
  2.14.1 Arabic MTE Research .................................... 96
2.14.1.1 MTE Study in Iraq .......................................................... 96
2.14.1.2 Very Recent Arabic MTE Studies ................................. 97
2.14.2 Review of Non-Arabic MTE Research ............................. 98
  2.14.2.1 Evaluations of Different MT Aspects and MT Output Quality .................................................. 99
  2.14.2.2 ISLE-Based Evaluations .............................................. 101
  2.14.2.3 Automatic Evaluations of MT Systems ....................... 102
2.15 Summary .............................................................................. 104
Notes to Chapter Two ................................................................ 105

CHAPTER THREE: THE MODEL AND REQUIREMENTS OF EVALUATING ARABIC MT SYSTEMS
3.1 Evaluation of MT Systems : Background .............................. 108
3.2 Levels of Evaluation ............................................................. 110
  3.2.1 Macroevaluation ............................................................. 110
  3.2.2 Microevaluation ............................................................. 111
3.3 The Framework Model of MTE ............................................ 114
3.3.1 Establishing Evaluation Requirements............................................. 114
  3.3.1.1 Functional and Non-Functional Requirements ..................... 118
  3.3.1.2 Quality : A Wider Outlook.............................................. 119
  3.3.1.3 Evaluation Criteria ..................................................... 123
    3.3.1.3.1 Functional Criteria ............................................ 124
    3.3.1.3.2 Non-Functional Criteria ................................. 128
3.3.2 Specifying the Evaluation ..................................................... 129
  3.3.2.1 Metrics ........................................................................ 132
    3.3.2.1.1 An Introduction to Metrics ................................. 132
    3.3.2.1.2 Validity of Measures .......................................... 133
    3.3.2.1.3 Reliability of Measures ....................................... 134
  3.3.2.2 Methods for System Measurement ................................. 134
  3.3.2.3 Description of Metrics and Criteria Scoring Methods
    in the Present Evaluation .................................................... 136
    3.3.2.3.1 Evaluation of Readability .................................... 137
    3.3.2.3.2 Evaluation of Fidelity ........................................ 138
    3.3.2.3.3 Evaluation of Terminology .................................. 140
    3.3.2.3.4 Evaluation of Syntax .......................................... 142
    3.3.2.3.5 Evaluation of Morphology ................................... 143
    3.3.2.3.6 Evaluation of Input-to-output Translation ............ 144
    3.3.2.3.7 Evaluation of Adaptability .................................. 145
    3.3.2.3.8 Evaluation of Storage ........................................ 146
    3.3.2.3.9 Evaluation of Flexibility .................................... 146
  3.3.2.4 GSQ Weighting Scheme .................................................... 147
3.3.3 Producing the Evaluation Plan ................................................. 150
3.3.4 Executing the Evaluation ....................................................... 150
3.4 Summary ................................................................................. 151
CHAPTER FOUR: EXPERIMENTAL DESIGN AND EXECUTION OF MT EVALUATION

4.1 Key Concepts in Evaluation ................................................................. 153

4.2 Experimental Design ........................................................................... 154

4.2.1 Text Sampling .................................................................................. 154

4.2.2 Three Systems Tested ...................................................................... 157

4.2.3 Evaluators ....................................................................................... 159

4.2.4 User and Task Description ............................................................... 160

4.2.5 A User-Oriented Model of Test Types ........................................... 162

4.2.5.1 Systematic Testing ...................................................................... 162

4.2.5.1.1 Task-Based MT Evaluation .................................................. 162

4.2.5.1.2 Benchmark Testing ............................................................... 163

4.2.5.2 Testing Instruments ..................................................................... 164

4.2.5.2.1 Checklists ............................................................................. 164

4.3 Evaluation Procedure .......................................................................... 165

4.3.1 Implementation of Evaluation Methodology .................................... 166

4.3.1.1 Testing of Functional Criteria .................................................... 166

4.3.1.1.1 Readability Rating ............................................................... 167

4.3.1.1.2 Fidelity Rating ..................................................................... 168

4.3.1.1.3 Domain Terminology Rating ............................................. 169

4.3.1.1.4 Syntax Rating ...................................................................... 170

4.3.1.1.5 Morphology Rating ............................................................ 171

4.3.1.2 Testing of Non-Functional Criteria ............................................ 172

4.3.1.2.1 Testing the Input-to-Output Translation ............................... 172

4.3.1.2.2 Testing Adaptability ............................................................. 173

4.3.1.2.3 Testing Storage .................................................................. 174

4.3.1.2.4 Testing Flexibility ............................................................... 177
CHAPTER FIVE: CLASSIFICATION AND DISCUSSION OF EVALUATION RESULTS

5.1 Analysis and Classification of Results ............................................ 187

5.1.1 Evaluation Results of Individual Software Criteria ................. 187
   5.1.1.1 Functional Evaluation Results ........................................ 187
      5.1.1.1.1 Variation in Scores between Raters ......................... 188
      5.1.1.1.2 Comparison of Systems for Text Types ............... 197
      5.1.1.1.3 Interrelation among Functional Criteria ............ 204
   5.1.1.2 Ranking of MT Systems for Non-Functional Criteria ..... 206
      5.1.1.2.1 Speed of Translation Results ............................... 206
      5.1.1.2.2 Adaptability Results ........................................ 210
      5.1.1.2.3 Storage Results ............................................. 210
      5.1.1.2.4 Flexibility Results ........................................ 212

5.1.2 Evaluation Results of GSQ .................................................... 214
   5.1.2.1 Systems’ Quality Assessment and Comparison for Criteria
      Types .............................................................................. 214
   5.1.2.2 Systems’ Quality Assessment and Comparison for Text
      Types .............................................................................. 217

5.2 Problems of MT Quality .............................................................. 220

   5.2.1 Cognitive Problems ............................................................. 221
      5.2.1.1 Grotesque Word Order and Choice ......................... 221
      5.2.1.2 Understandable Clumsy Translation ....................... 225
      5.2.1.3 Unknown and Mistranslated Critical Words .......... 227
      5.2.1.4 Partially Correct Translation ................................. 228
      5.2.1.5 Adherence to SL Structure .................................... 229
APPENDICES ................................................................. 324
### KEY TO ABBREVIATIONS AND ACRONYMS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>ALPAC</td>
<td>Automatic Language Processing Advisory Committee</td>
</tr>
<tr>
<td>AMTA</td>
<td>Association for MT in the Americas</td>
</tr>
<tr>
<td>ARAMED</td>
<td>European Group for Translating Medical Texts</td>
</tr>
<tr>
<td>ARPA</td>
<td>American Research Programme Association</td>
</tr>
<tr>
<td>ATA</td>
<td>Arabic Trade Agreement</td>
</tr>
<tr>
<td>CAT</td>
<td>Computer-Aided Translation</td>
</tr>
<tr>
<td>CL</td>
<td>Computational Linguistics</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processor Unit</td>
</tr>
<tr>
<td>DARPA</td>
<td>Defense Advanced Research Projects Agency</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>EAGLES</td>
<td>Expert Advisory Group on Language Engineering Standards</td>
</tr>
<tr>
<td>Eval.</td>
<td>Evaluator</td>
</tr>
<tr>
<td>Evals.</td>
<td>Evaluators</td>
</tr>
<tr>
<td>FAHQT</td>
<td>Fully Automatic High Quality Translation</td>
</tr>
<tr>
<td>FEMTI</td>
<td>Framework for the Evaluation of Machine Translation in the ISLE</td>
</tr>
<tr>
<td>GSQ</td>
<td>General Software Quality</td>
</tr>
<tr>
<td>GW</td>
<td>Golden AL-Wafi</td>
</tr>
<tr>
<td>HLT</td>
<td>Human Language Technology</td>
</tr>
<tr>
<td>HT</td>
<td>Human Translation</td>
</tr>
<tr>
<td>IBM</td>
<td>International Business Machine</td>
</tr>
<tr>
<td>IEC</td>
<td>International Electrotechnical Commission</td>
</tr>
<tr>
<td>ISLE</td>
<td>International Standards for Language Engineering</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ISSCO</td>
<td>Institute for Semantic and Cognitive Studies</td>
</tr>
<tr>
<td>JEDIA</td>
<td>Japan Electronic Industry Development Association</td>
</tr>
<tr>
<td>LREC</td>
<td>Language Requirements Evaluation Conference</td>
</tr>
</tbody>
</table>
UNKOWN ABBREVIATIONS AND ACRONYMS

BLUE  (An Automatic MTE Method)

COBALT (An Evaluation Research Project)

LUNAR  (An MT System)

KWIC   (indexes)

OVUM   (An MTE Project)
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. The Various Sample Texts in the Present Evaluation each with the Number of Sentences Selected</td>
<td>157</td>
</tr>
<tr>
<td>4.2. Hardware and Software Features of the PC Used in the Present Evaluation</td>
<td>157</td>
</tr>
<tr>
<td>4.3. The Number of Words, Expressions and Terms Inserted by the Researcher in each Dictionary of the MT Systems</td>
<td>158</td>
</tr>
<tr>
<td>4.4. Sources and Costs of MT Systems under Evaluation</td>
<td>159</td>
</tr>
<tr>
<td>4.5. Identification of the MT Systems</td>
<td>159</td>
</tr>
<tr>
<td>5.1. Two Evaluators’ Scoring Results of Readability, Fidelity and Syntax</td>
<td>190</td>
</tr>
</tbody>
</table>
5.2. Correlation Coefficient between the Results of Evaluator 1 and Evaluator 2 ................................................................. 196
5.3. Correlation Coefficient Matrix of the Functional Criteria of Golden Al-Wafi ................................................................. 204
5.4. Correlation Coefficient Matrix of the Functional Criteria of Al-Mutarjim Al-Arabey ......................................................... 204
5.5. Correlation Coefficient Matrix of the Functional Criteria of An-Nakel Al-Arabi ................................................................. 204
5.6. Speed of Translation with Respect to Text Size ........................................ 207
5.7. Translation Time for the Selected Texts ............................................ 207
5.8. Speed of Translation and the Evaluation Speed % of the Selected Texts ............................................................................. 208
5.9. The Effect of Processor Speed on the Translation Speed .................. 209
5.10. The Fastest MT System in Translating Typical Text of 29 Kb .......... 210
5.11. Average Values of Adaptability Based on the Checklist Results for the Arabic MT Systems .................................................... 210
5.12. Direct Translation in the Background for GW and MA .................... 211
5.13. Average Values of Storage Based on the Results of the Two Evaluators for the Arabic MT Systems ........................................... 212
5.14. Lexical Flexibility of Arabic MT Systems ........................................ 213
5.15. Average Values of Flexibility Based on the Checklist Results for the Arabic MT Systems ................................................... 213
5.16. General Software Quality Based on Functional and Non-Functional Criteria ...................................................................... 215
5.17. Average Values of Mother-Node Criteria in Terms of Systems Functional Performance ....................................................... 217
5.18. GSQ Based on Functional and Non-Functional Total Performance ........................................................................217
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 2.1 General Diagram Showing How AL-Mutarjim AL-Arabey Runs</td>
<td>66</td>
</tr>
<tr>
<td>Fig. 2.2 The Tree-Structure of the Sentence</td>
<td>67</td>
</tr>
<tr>
<td>Fig. 2.3 The Tree-Structure of the Relationship between the English Text and its Arabic Translation</td>
<td>68</td>
</tr>
<tr>
<td>Fig. 2.4 The Whole Translation Process in AL-Mutarjim AL-Arabey When the Verb is Intransitive/Simple Past Tense</td>
<td>69</td>
</tr>
<tr>
<td>Fig. 2.5 The Main Window (and Standard Toolbar) of Al-Mutarjim AL-Arabey 3.00</td>
<td>73</td>
</tr>
<tr>
<td>Fig. 2.6 The Setting Window of Al-Mutajim AL-Arabey 3.00</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 2.7 Translation of Proper Nouns and Abbreviations</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 2.8 Diacritics on the Arabic Words</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 2.9 Alternative Words in Al-Mutarjim AL-Arabey 3.00</td>
<td>74</td>
</tr>
<tr>
<td>Fig. 2.10 Translation of Words According to the Theme of the Text</td>
<td>75</td>
</tr>
<tr>
<td>Fig. 2.11 Multiple Meanings of Words in Al-Mutarjim AL-Arabey 3.00</td>
<td>75</td>
</tr>
<tr>
<td>Fig. 2.12 A General Overview of An-Nakel AL-Arabi</td>
<td>83</td>
</tr>
<tr>
<td>Fig. 2.13 The Main Window (and Standard Toolbar) of An-Nakel Al-Arabi 2.00</td>
<td>84</td>
</tr>
<tr>
<td>Fig. 2.14 Document Commands of An-Nakel Al-Arabi</td>
<td>84</td>
</tr>
</tbody>
</table>
Fig. 2.15 Translation Menu Commands ............................................................ 84
Fig. 2.16 Dictionary Menu Commands ............................................................. 85
Fig. 2.17 Option Menu Commands ................................................................. 85
Fig. 2.18 Alternative Meanings for Words and Idioms in An-Nakel Al-Arabi ................................................................. 85
Fig. 2.19 Interactive Translation in An-Nakel Al-Arabi ................................. 86
Fig. 2.20 Main Window (and Standard Toolbar) of Golden Al-Wafi ............. 88
Fig. 2.21 The Setting Window of Golden Al-Wafi ........................................... 89
Fig. 2.22 Suggestions for Spelling Checker ..................................................... 89
Fig. 2.23 Multiple Meanings of Word ............................................................. 89
Fig. 2.24 Alternative Meanings of Word ......................................................... 89
Fig. 2.25 Diacritics on Arabic Words ............................................................... 89
Fig. 2.26 Translation and Transliteration of Proper Nouns and Abbreviations ................................................................................. 90
Fig. 3.1 V-diagram Showing the Place of Evaluation in Software Development ......................................................................................... 108
Fig. 3.2 Macro-and Micro-Evaluations of MT Systems ................................ 113
Fig. 3.3 The EAGLES 7 Major Steps for a Successful Evaluation ............... 115
Fig. 3.4 Stages in the Evaluation Process ....................................................... 116
Fig. 3.5 The Evaluation Requirements .......................................................... 116
Fig. 3.6 Mapping Categories of Requirements to Evaluation Criteria .......... 119
Fig. 3.7 Software Quality Characteristics .................................................... 121
Fig. 3.8 Relationship between Internal Quality, External Quality and Quality in Use ................................................................................... 121
Fig. 3.9 Approaches to Software Product Quality ........................................ 122
Fig. 3.10 Quality in the Software Lifecycle .................................................. 123
Fig. 3.11 Model of Evaluating the Total Performance (the External Quality of Translation) of the Arabic MT Systems ........................................ 125
Fig. 4.1 Factors Determining Evaluation Design ........................................... 155
Fig. 4.2 Direct Translation Dialogue Box in AL-Mutarjim AL-Arabey 3.00 .......................................................... 176
Fig. 4.3 Direct Translation Dialogue Box in Golden AL-Wafi 1.00 .............. 176
Fig. 4.4 Bi-lingual Dictionary of Golden AL-Wafi 1.00 ......................... 177
Fig. 4.5 Specialized Dictionaries of Golden AL-Wafi 1.00 ...................... 178
Fig. 4.6 Bi-lingual Dictionary of AL-Mutarjim AL-Arabey 3.00 .......... 178
Fig. 4.7 Specialized Dictionaries of AL-Mutarjim AL-Arabey 3.00 .......... 179
Fig. 4.8 User and Abbreviations Dictionaries of AL-Mutarjim AL-Arabey 3.00 .......................................................... 180
Fig. 4.9 Preference Dictionary of AL-Mutarjim AL-Arabey 3.00 .......... 180
Fig. 4.10 Translation Memory in An -Nakel AL-Arabi ......................... 181
Fig. 4.11 User Dictionary of An-Nakel AL-Arabi ................................ 182
Fig. 4.12 Idioms Dictionary of An-Nakel AL-Arabi ................................ 184
Fig. 5.1 Variation of the Three MT Systems’ Readability Based on Text Types .......................................................... 198
Fig. 5.2 Variation of the Three MT Systems’ Fidelity Based on Text Types ........................................................................ 199
Fig. 5.3 Variation of the Three MT Systems’ Terminology Based on Text Types ........................................................................ 200
Fig. 5.4 Variation of the Three MT Systems’ Syntax Based on Text Types ................................................................. 202
Fig. 5.5 Variation of the Three MT Systems’ Morphology Based on Text Type .............................................................. 203
Fig. 5.6 Average Values of Functional Criteria for Arabic MT Systems ........................................................................ 203
Fig. 5.7 Scatter Plot of the Readability and Terminology Results of MA System .............................................................. 206
Fig. 5.8 Average Values of Non-Functional Criteria Showing the Range of Arabic MT Systems ........................................ 214
Fig. 5.9 The Percentages of the Total Quality for the Three Arabic MT Systems ............................................................ 216
Fig. 5.10 The Variation of Total Quality for the Three MT Systems Based on Text Types ................................................ 218
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Subject</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAPTER ONE: INTRODUCTION</td>
<td></td>
</tr>
<tr>
<td>1.1 Preliminaries</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Evaluation of Machine Translation Systems</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Evaluation of Existing Arabic Systems</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Problem</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Hypotheses</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Aims</td>
<td>9</td>
</tr>
<tr>
<td>1.5 Scope</td>
<td>10</td>
</tr>
<tr>
<td>1.6 Procedure</td>
<td>11</td>
</tr>
<tr>
<td>1.7 Significance</td>
<td>13</td>
</tr>
<tr>
<td>1.8 Plan</td>
<td>15</td>
</tr>
<tr>
<td>CHAPTER TWO: EVALUATING MT SYSTEMS: SOME BASIC ISSUES</td>
<td></td>
</tr>
<tr>
<td>2.1 General Overview</td>
<td>20</td>
</tr>
<tr>
<td>2.2 Scope of Evaluation</td>
<td>23</td>
</tr>
<tr>
<td>2.3 The Role of Evaluation in MT</td>
<td>26</td>
</tr>
<tr>
<td>2.4 Some History</td>
<td>27</td>
</tr>
<tr>
<td>2.4.1 Background</td>
<td>27</td>
</tr>
<tr>
<td>2.4.2 Some Past Evaluations</td>
<td>30</td>
</tr>
<tr>
<td>2.4.2.1 ALPAC</td>
<td>30</td>
</tr>
<tr>
<td>2.4.2.2 (D) ARPA</td>
<td>32</td>
</tr>
</tbody>
</table>
2.4.2.3 The Hewlett Packard Laboratories Study .................. 33
2.4.2.4 MUC-3 ..................................................................... 35
2.4.2.5 COBALT .................................................................. 36

2.5 Standards Work Related to Evaluation ................................. 37
2.5.1 Previous Approaches to MTE ........................................ 40
2.5.1.1 JEIDA (1992) .......................................................... 40
2.5.1.2 The EAGLES Evaluation Guidelines .......................... 41
2.5.1.3 The EAGLES 7-Step Recipe ................................. 44
2.5.1.4 The ISO/IEC Standards for Software Evaluation .......... 47
   2.5.1.4.1 A Growing Set of Standards ................................. 47
   2.5.1.4.2 The Definition of a Quality Model ....................... 47
   2.5.1.4.3 Stages in the Evaluation Process ....................... 49
2.5.1.5 The ISLE MTE Framework ........................................ 51
   2.5.1.5.1 Origins and First Version ................................... 52
   2.5.1.5.2 Key Principle of the Second Draft ....................... 53
   2.5.1.5.3 A Framework for the Evaluation of Machine
   Translation in ISLE (2003) ........................................ 54
   2.5.1.5.4 Analyzing the Behaviour of Measures .................. 56

2.6 New Trends and Future Work in MTE .................................. 58
2.7 Formal and Functional Approaches to MTE ............................ 61
2.8 Evaluation for MT Stakeholders ........................................ 63
2.9 Types and Purposes of MTE .............................................. 66
   2.9.1 Feasibility Evaluation ............................................. 66
   2.9.2 Internal Evaluation ................................................ 67
   2.9.3 Black-box Vs. Glass-box Evaluation ........................... 68
   2.9.4 Diagnostic Evaluation ............................................ 69
   2.9.5 Declarative Evaluation ............................................ 69
2.9.6. Adequacy Evaluation ................................................................. 70
2.9.7 Performance Evaluation ............................................................. 71
2.9.8 Operational Evaluation ............................................................... 71
2.9.9 Usability Evaluation ................................................................. 72
2.9.10 Comparison Evaluation ............................................................. 72
2.10 The Object of Evaluation ............................................................. 73
2.11 MT Systems .................................................................................. 75
   2.11.1 Background ............................................................................ 75
   2.11.2 MT Systems and Text Type Considerations ............................ 77
   2.11.3 Classification of MT Systems .................................................. 79
      2.11.3.1 Typology of MT Systems by Text Type ....................... 79
      2.11.3.2 Typology of MT Systems by Degree of Automation ... 79
   2.11.4 Description of Arabic MT Systems ........................................ 80
      2.11.4.1 Sakhar ............................................................................ 82
      2.11.4.2 Turjman ........................................................................ 83
      2.11.4.3 ArabTrans ..................................................................... 83
      2.11.4.4 AL-Wafi ........................................................................ 84
      2.11.4.5 SYSTRAN ..................................................................... 87
      2.11.4.6 Applied Technology (Apptek) ....................................... 87
   2.11.4.7 Other Arabic MT Systems and Projects ......................... 89
      2.11.4.7.1 Electronic Research Institute: The National Research Council in Cairo .......... 89
      2.11.4.7.2 The Scientific Center in Cairo (IBM) ....................... 89
      2.11.4.7.3 XEROX Company ..................................................... 90
      2.11.4.7.4 Luts Company ......................................................... 90
      2.11.4.7.5 The House of Computerizing the Arabic Text .......... 90
2.11.4.7.6 Microsoft.................................................90
2.11.4.8 Arabic MT Systems under Evaluation.........................90
  2.11.4.8.1 AL-Mutarjim AL-Arabey...............................91
    2.11.4.8.1.1 General ........................................91
    2.11.4.8.1.2 The General Dictionary ......................94
    2.11.4.8.1.3 How Does AL-Mutarjim AL-Arabey Work? .........95
    2.11.4.8.1.4 General Issues............................98
    2.11.4.8.1.5 AL-Mutarjim AL-Arabey 3.00 (2002) ..............100
  2.11.4.8.2 An-Nakel AL-Arabi..................................106
    2.11.4.8.2.1 General ........................................106
    2.11.4.8.2.2 How It Works? .............................108
    2.11.4.8.2.3 Levels of System Operation ........................110
  2.11.4.8.3 Golden Al-Wafi 1.00 (2002) ..........................121

2.12 The State of Arabic Machine Translation..........................126
  2.12.1 Existing Arabic MT Systems and Arabic MTE...............127
2.13 Arabic MT and the Internet........................................129
2.14 Recent Research on MTE .............................................133
  2.14.1 Arabic MTE Research.........................................134
    2.14.1.1 MTE Study in Iraq ..................................134
    2.14.1.2 Very Recent Arabic MTE Studies......................135
  2.14.2 Review of Non-Arabic MTE Research............................137
    2.14.2.1 Evaluations of Different MT Aspects and MT Output Quality ...........................................138
    2.14.2.2 ISLE-Based Evaluations ................................141
    2.14.2.3 Automatic Evaluations of MT Systems .................142
xxxvi

2.15 Summary .............................................................................................................. 145

Notes to Chapter Two .................................................................................................. 146

CHAPTER THREE: THE MODEL AND REQUIREMENTS OF EVALUATING ARABIC MT SYSTEMS

3.1 Evaluation of MT Systems: Background ............................................................... 152

3.2 Levels of Evaluation ........................................................................................... 155

3.2.1 Macroevaluation ............................................................................................ 156

3.2.2 Microevaluation ............................................................................................. 157

3.3 The Framework Model of MTE ......................................................................... 161
3.3.1 Establishing Evaluation Requirements ........................................... 161
  3.3.1.1 Functional and Non-Functional Requirements .................. 166
  3.3.1.2 Quality : A Wider Outlook ........................................... 168
  3.3.1.3 Evaluation Criteria .................................................... 173
    3.3.1.3.1 Functional Criteria ........................................... 175
    3.3.1.3.2 Non-Functional Criteria .................................. 179
3.3.2 Specifying the Evaluation ...................................................... 181
  3.3.2.1 Metrics ....................................................................... 184
    3.3.2.1.1 An Introduction to Metrics ................................. 184
    3.3.2.1.2 Validity of Measures ......................................... 186
    3.3.2.1.3 Reliability of Measures ...................................... 187
  3.3.2.2 Methods for System Measurement .................................. 188
  3.3.2.3 Description of Metrics and Criteria Scoring Methods
    in the Present Evaluation .................................................... 190
    3.3.2.3.1 Evaluation of Readability .................................. 191
    3.3.2.3.2 Evaluation of Fidelity ......................................... 193
    3.3.2.3.3 Evaluation of Terminology .................................. 196
    3.3.2.3.4 Evaluation of Syntax .......................................... 198
    3.3.2.3.5 Evaluation of Morphology .................................. 200
    3.3.2.3.6 Evaluation of Input-to-output Translation ............ 201
    3.3.2.3.7 Evaluation of Adaptability ................................... 203
    3.3.2.3.8 Evaluation of Storage ......................................... 204
    3.3.2.3.9 Evaluation of Flexibility .................................... 205
  3.3.2.4 GSQ Weighting Scheme ..................................................... 206
3.3.3 Producing the Evaluation Plan .................................................. 210
3.3.4 Executing the Evaluation .......................................................... 210
3.4 Summary ......................................................................................... 210
CHAPTER FOUR: EXPERIMENTAL DESIGN AND EXECUTION OF MT EVALUATION

4.1 Key Concepts in Evaluation ................................................................. 215

4.2 Experimental Design ........................................................................... 217
  4.2.1 Text Sampling ............................................................................. 218
  4.2.2 Three Systems Tested ................................................................. 220
  4.2.3 Evaluators .................................................................................. 223
  4.2.4 User and Task Description ......................................................... 225
  4.2.5 A User-Oriented Model of Test Types ....................................... 228
    4.2.5.1 Systematic Testing ............................................................... 228
      4.2.5.1.1 Task-Based MT Evaluation ......................................... 229
      4.2.5.1.2 Benchmark Testing ...................................................... 230
    4.2.5.2 Testing Instruments .............................................................. 231
      4.2.5.2.1 Checklists .................................................................... 232

4.3 Evaluation Procedure .......................................................................... 232
  4.3.1 Implementation of Evaluation Methodology .................................. 233
    4.3.1.1 Testing of Functional Criteria ............................................ 234
      4.3.1.1.1 Readability Rating ......................................................... 235
      4.3.1.1.2 Fidelity Rating ............................................................... 237
      4.3.1.1.3 Domain Terminology Rating ....................................... 238
      4.3.1.1.4 Syntax Rating ............................................................... 240
      4.3.1.1.5 Morphology Rating ...................................................... 241
    4.3.1.2 Testing of Non-Functional Criteria ..................................... 242
      4.3.1.2.1 Testing the Input-to-Output Translation ...................... 243
      4.3.1.2.2 Testing Adaptability .................................................... 244
      4.3.1.2.3 Testing Storage ............................................................ 245
      4.3.1.2.4 Testing Flexibility ......................................................... 247
4.4 Summary ......................................................................................................................... 257
Notes to Chapter Four ........................................................................................................ 259

CHAPTER FIVE: CLASSIFICATION AND DISCUSSION OF EVALUATION RESULTS

5.1 Analysis and Classification of Results ................................................................. 262

5.1.1 Evaluation Results of Individual Software Criteria ......................... 263

5.1.1.1 Functional Evaluation Results ................................................................. 263
5.1.1.1.1 Variation in Scores between Raters .................................................. 263
5.1.1.1.2 Comparison of Systems for Text Types ...................................... 276
5.1.1.1.3 Interrelation among Functional Criteria .................................... 285

5.1.1.2 Ranking of MT Systems for Non-Functional Criteria ............ 288

5.1.1.2.1 Speed of Translation Results ......................................................... 289
5.1.1.2.2 Adaptability Results ..................................................................... 293
5.1.1.2.3 Storage Results .......................................................................... 295
5.1.1.2.4 Flexibility Results ...................................................................... 297

5.1.2 Evaluation Results of GSQ .............................................................. 300

5.1.2.1 Systems’ Quality Assessment and Comparison for Criteria Types .......................................................................................................................... 300
5.1.2.2 Systems’ Quality Assessment and Comparison for Text Types .......................................................................................................................... 304

5.2 Problems of MT Quality ................................................................. 308

5.2.1 Cognitive Problems ................................................................................. 309

5.2.1.1 Grotesque Word Order and Choice ............................................ 309
5.2.1.2 Understandable Clumsy Translation ............................................ 315
5.2.1.3 Unknown and Mistranslated Critical Words ................................ 318
5.2.1.4 Partially Correct Translation ......................................................... 320
5.2.1.5 Adherence to SL Structure .............................................................. 322
5.2.1.6 Loss and Distortion of Information ........................................ 326
5.2.2 Linguistic Problems ...................................................................... 338
   5.2.2.1 Inadequate Terminology Interpretation ................................. 338
   5.2.2.2 Incorrect Rendition of SL Grammatical Structure ............... 356
   5.2.2.3 Erroneously Recognised TL Morphology ......................... 380
5.2.3 Operational Problems .................................................................. 396
   5.2.3.1 Impediments in Measuring Speed ........................................ 396
   5.2.3.2 Limitations of User Dictionaries ........................................... 398

5.3 Summary ............................................................................................. 408
Notes to Chapter Five ........................................................................... 409

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS
6.1 Current State of Arabic MT Systems ............................................. 412
   6.1.1 General Conclusion ...................................................................... 412
   6.1.2 Strengths and Weaknesses of the MT Systems Evaluated .......... 413
   6.1.3 Other Findings of the Present Evaluation ..................................... 416
6.2 Some Sources of Arabic MT and MTE Problems ............................ 424
   6.2.1 Variation between SL and TL Linguistic Structures ................. 424
   6.2.2 Deficiencies of Systems’ Dictionaries and Components ............. 425
   6.2.3 The Frail Nature of MT ................................................................. 426
   6.2.4 Lacking Awareness of the Strategic Nature of Arabic MT ......... 426
   6.2.5 Lacking Awareness of the Significance of MTE ........................... 427
6.3 Recommendations ............................................................................. 428
6.4 Suggestions for Further Research .................................................... 432

REFERENCES .......................................................................................... 433
APPENDICES ........................................................................................................461
# KEY TO ABBREVIATIONS AND ACRONYMS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>ALPAC</td>
<td>Automatic Language Processing Advisory Committee</td>
</tr>
<tr>
<td>AMTA</td>
<td>Association for MT in the Americas</td>
</tr>
<tr>
<td>ARAMED</td>
<td>European Group for Translating Medical Texts</td>
</tr>
<tr>
<td>ARPA</td>
<td>American Research Programme Association</td>
</tr>
<tr>
<td>ATA</td>
<td>Arabic Trade Agreement</td>
</tr>
<tr>
<td>CAT</td>
<td>Computer-Aided Translation</td>
</tr>
<tr>
<td>CL</td>
<td>Computational Linguistics</td>
</tr>
<tr>
<td>CPU</td>
<td>Central Processor Unit</td>
</tr>
<tr>
<td>DARPA</td>
<td>Defense Advanced Research Projects Agency</td>
</tr>
<tr>
<td>DT</td>
<td>Decision Tree</td>
</tr>
<tr>
<td>EAGLES</td>
<td>Expert Advisory Group on Language Engineering Standards</td>
</tr>
<tr>
<td>Eval.</td>
<td>Evaluator</td>
</tr>
<tr>
<td>Evals.</td>
<td>Evaluators</td>
</tr>
<tr>
<td>FAHQT</td>
<td>Fully Automatic High Quality Translation</td>
</tr>
<tr>
<td>FEMTI</td>
<td>Framework for the Evaluation of Machine Translation in the ISLE</td>
</tr>
<tr>
<td>GSQ</td>
<td>General Software Quality</td>
</tr>
<tr>
<td>GW</td>
<td>Golden AL-Wafi</td>
</tr>
<tr>
<td>HLT</td>
<td>Human Language Technology</td>
</tr>
<tr>
<td>HT</td>
<td>Human Translation</td>
</tr>
<tr>
<td>IBM</td>
<td>International Business Machine</td>
</tr>
<tr>
<td>IEC</td>
<td>International Electrotechnical Commission</td>
</tr>
<tr>
<td>ISLE</td>
<td>International Standards for Language Engineering</td>
</tr>
<tr>
<td>ISO</td>
<td>International Organization for Standardization</td>
</tr>
<tr>
<td>ISSCO</td>
<td>Institute for Semantic and Cognitive Studies</td>
</tr>
<tr>
<td>JEDIA</td>
<td>Japan Electronic Industry Development Association</td>
</tr>
<tr>
<td>LREC</td>
<td>Language Requirements Evaluation Conference</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>MA</td>
<td>Al-Mutarjim Al-Arabey</td>
</tr>
<tr>
<td>MAT</td>
<td>Machine-Aided Translation</td>
</tr>
<tr>
<td>MAHT</td>
<td>Machine –Aided Human Translation</td>
</tr>
<tr>
<td>ME</td>
<td>Millennium</td>
</tr>
<tr>
<td>MLTS</td>
<td>Multi-Lingual Translation Service</td>
</tr>
<tr>
<td>MT</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>MTE</td>
<td>Machine Translation Evaluation</td>
</tr>
<tr>
<td>MUC</td>
<td>Message Understanding Conference</td>
</tr>
<tr>
<td>NA</td>
<td>An-Nakel Al-Arabi</td>
</tr>
<tr>
<td>NAACL</td>
<td>North American Association for Computational Linguistics</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>OCR</td>
<td>Optical Character Recognition</td>
</tr>
<tr>
<td>PC</td>
<td>Personal Computer</td>
</tr>
<tr>
<td>R &amp; D</td>
<td>Research and Development</td>
</tr>
<tr>
<td>SET</td>
<td>Sakhr Enterprise Translation System</td>
</tr>
<tr>
<td>SL</td>
<td>Source Language</td>
</tr>
<tr>
<td>ST</td>
<td>Source Text</td>
</tr>
<tr>
<td>TEMA A</td>
<td>A Tested Study of Evaluation Methodology: Authoring Aids</td>
</tr>
<tr>
<td>TL</td>
<td>Target Language</td>
</tr>
<tr>
<td>TM</td>
<td>Translation Memory</td>
</tr>
<tr>
<td>TT</td>
<td>Target Text</td>
</tr>
<tr>
<td>UAE</td>
<td>United Arab Emirates</td>
</tr>
<tr>
<td>UMIST</td>
<td>University of Manchester Institute of Translation Studies</td>
</tr>
<tr>
<td>UNDP</td>
<td>United Nations Development Programme</td>
</tr>
<tr>
<td>URL</td>
<td>Uniform Resources Location</td>
</tr>
</tbody>
</table>
UNKNOWN ABBREVIATIONS AND ACRONYMS

BLUE (An Automatic MTE Method)

COBALT (An Evaluation Research Project)

LUNAR (An MT System)

KWIC (indexes)

OVUM (An MTE Project)
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1. The Various Sample Texts in the Present Evaluation each with the Number of Sentences Selected</td>
<td>221</td>
</tr>
<tr>
<td>4.2. Hardware and Software Features of the PC Used in the Present Evaluation</td>
<td>221</td>
</tr>
<tr>
<td>4.3. The Number of Words, Expressions and Terms Inserted by the Researcher in each Dictionary of the MT Systems</td>
<td>223</td>
</tr>
<tr>
<td>4.4. Sources and Costs of MT Systems under Evaluation</td>
<td>224</td>
</tr>
<tr>
<td>4.5. Identification of the MT Systems</td>
<td>224</td>
</tr>
<tr>
<td>5.1. Two Evaluators’ Scoring Results of Readability, Fidelity and Syntax</td>
<td>266</td>
</tr>
<tr>
<td>5.2. Correlation Coefficient between the Results of Evaluator 1 and Evaluator 2</td>
<td>276</td>
</tr>
<tr>
<td>5.3. Correlation Coefficient Matrix of the Functional Criteria of Golden Al- Wafi</td>
<td>296</td>
</tr>
<tr>
<td>5.4. Correlation Coefficient Matrix of the Functional Criteria of Al-Mutarjim Al-Arabey</td>
<td>296</td>
</tr>
</tbody>
</table>
5.5. Correlation Coefficient Matrix of the Functional Criteria of An-Nakel Al-Arabi ................................................................. 296
5.6. Speed of Translation with Respect to Text Size ............................................. 289
5.7. Translation Time for the Selected Texts .......................................................... 291
5.8. Speed of Translation and the Evaluation Speed % of the Selected Texts ................................................................. 292
5.9. The Effect of Processor Speed on the Translation Speed......................... 293
5.10. The Fastest MT System in Translating Typical Text of 29 Kb ............ 293
5.11. Average Values of Adaptability Based on the Checklist Results for the Arabic MT Systems ................................................................. 294
5.12. Direct Translation in the Background for GW and MA ...................... 296
5.13. Average Values of Storage Based on the Results of the Two Evaluators for the Arabic MT Systems ................................................................. 297
5.14. Lexical Flexibility of Arabic MT Systems .............................................. 298
5.15. Average Values of Flexibility Based on the Checklist Results for the Arabic MT Systems ................................................................. 299
5.16. General Software Quality Based on Functional and Non-Functional Criteria ................................................................. 301
5.17. Average Values of Mother-Node Criteria in Terms of Systems Functional Performance ................................................................. 303
5.18. GSQ Based on Functional and Non-Functional Total Performance ................................................................. 304
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig. 2.1 General Diagram Showing How AL-Mutarjim AL-Arabey Runs</td>
<td>93</td>
</tr>
<tr>
<td>Fig. 2.2 The Tree-Structure of the Sentence</td>
<td>95</td>
</tr>
<tr>
<td>Fig. 2.3 The Tree-Structure of the Relationship between the English Text and its Arabic Translation</td>
<td>96</td>
</tr>
<tr>
<td>Fig. 2.4 The Whole Translation Process in AL-Mutarjim AL-Arabey When the Verb is Intransitive/Simple Past Tense</td>
<td>97</td>
</tr>
<tr>
<td>Fig. 2.5 The Main Window (and Standard Toolbar) of AL-Mutarjim AL-Arabey 3.00</td>
<td>103</td>
</tr>
<tr>
<td>Fig. 2.6 The Setting Window of AL-Mutarjim AL-Arabey 3.00</td>
<td>104</td>
</tr>
<tr>
<td>Fig. 2.7 Translation of Proper Nouns and Abbreviations</td>
<td>104</td>
</tr>
<tr>
<td>Fig. 2.8 Diacritics on the Arabic Words</td>
<td>104</td>
</tr>
<tr>
<td>Fig. 2.9 Alternative Words in AL-Mutarjim AL-Arabey 3.00</td>
<td>105</td>
</tr>
<tr>
<td>Fig. 2.10 Translation of Words According to the Theme of the Text</td>
<td>105</td>
</tr>
<tr>
<td>Fig. 2.11 Multiple Meanings of Words in AL-Mutarjim AL-Arabey 3.00</td>
<td>106</td>
</tr>
<tr>
<td>Fig. 2.12 A General Overview of An-Nakel AL-Arabi</td>
<td>117</td>
</tr>
<tr>
<td>Fig. 2.13 The Main Window (and Standard Toolbar) of An-Nakel Al-Arabi 2.00</td>
<td>118</td>
</tr>
</tbody>
</table>
Fig. 2.14 Document Commands of An-Nakel Al-Arabi ......................... 118
Fig. 2.15 Translation Menu Commands .................................................. 119
Fig. 2.16 Dictionary Menu Commands ...................................................... 119
Fig. 2.17 Option Menu Commands .......................................................... 119
Fig. 2.18 Alternative Meanings for Words and Idioms in An-Nakel Al-Arabi .......................................................... 120
Fig. 2.19 Interactive Translation in An-Nakel Al-Arabi ................................ 120
Fig. 2.20 Main Window (and Standard Toolbar) of Golden Al-Wafi ............................ 123
Fig. 2.21 The Setting Window of Golden Al-Wafi .................................... 123
Fig. 2.22 Suggestions for Spelling Checker ................................................ 124
Fig. 2.23 Multiple Meanings of Word .......................................................... 124
Fig. 2.24 Alternative Meanings of Word ...................................................... 125
Fig. 2.25 Diacritics on Arabic Words .......................................................... 125
Fig. 2.26 Translation and Transliteration of Proper Nouns and Abbreviations .......................................................... 125
Fig. 3.1 V-diagram Showing the Place of Evaluation in Software Development .......................................................... 153
Fig. 3.2 Macro- and Micro-Evaluations of MT Systems .................................. 160
Fig. 3.3 The EAGLES 7 Major Steps for a Successful Evaluation ............... 162
Fig. 3.4 Stages in the Evaluation Process ..................................................... 163
Fig. 3.5 The Evaluation Requirements ........................................................ 164
Fig. 3.6 Mapping Categories of Requirements to Evaluation Criteria ........... 167
Fig. 3.7 Software Quality Characteristics ................................................... 170
Fig. 3.8 Relationship between Internal Quality, External Quality and Quality in Use ........................................................................................................171
Fig. 3.9 Approaches to Software Product Quality ..........................................172
Fig. 3.10 Quality in the Software Lifecycle ....................................................174
Fig. 4.1 Factors Determining Evaluation Design ............................................218
Fig. 4.2 Direct Translation Dialogue Box in AL-Mutarjim AL-Arabey 3.00 ..................................................................................................................248
Fig. 4.3 Direct Translation Dialogue Box in Golden AL-Wafi 1.00 ...............248
Fig. 4.4 Bi-lingual Dictionary of Golden AL-Wafi 1.00 .................................249
Fig. 4.5 Specialized Dictionaries of Golden AL-Wafi 1.00 ............................249
Fig. 4.6 Bi-lingual Dictionary of AL-Mutarjim AL-Arabey 3.00 ....................250
Fig. 4.7 Specialized Dictionaries of AL-Mutarjim AL-Arabey 3.00 ...............250
Fig. 4.8 User and Abbreviations Dictionaries of AL-Mutarjim AL-Arabey 3.00 ........................................................................................................252
Fig. 4.9 Preference Dictionary of AL-Mutarjim AL-Arabey 3.00 .................252
Fig. 4.10 Translation Memory in An-Nakel AL-Arabi ..................................253
Fig. 4.11 User Dictionary of An-Nakel AL-Arabi ..........................................254
Fig. 4.12 Idioms Dictionary of An-Nakel AL-Arabi ......................................256
Fig. 5.1 Variation of the Three MT Systems’ Readability Based on Text Types ........................................................................................................277
Fig. 5.2 Variation of the Three MT Systems’ Fidelity Based on Text Types ........................................................................................................279
Fig. 5.3 Variation of the Three MT Systems’ Terminology Based on Text Types ........................................................................................................280
Fig. 5.4 Variation of the Three MT Systems’ Syntax Based on Text Types ................................................................. 282
Fig. 5.5 Variation of the Three MT Systems’ Morphology Based on Text Type ................................................................. 284
Fig. 5.6 Average Values of Functional Criteria for Arabic MT Systems ............................................................................ 285
Fig. 5.7 Scatter Plot of the Readability and Terminology Results of MA System ................................................................. 288
Fig. 5.8 Average Values of Non-Functional Criteria Showing the Range of Arabic MT Systems ................................................. 299
Fig. 5.9 The Percentages of the Total Quality for the Three Arabic MT Systems ................................................................. 302
Fig. 5.10 The Variation of Total Quality for the Three MT Systems Based on Text Types ................................................................. 305
CHAPTER ONE
INTRODUCTION

1.1 Preliminaries

1.1.1 Evaluation of Machine Translation Systems

Computers are information processing systems. In computers, we recognize a new kind of intelligence different from ours, in some way more powerful and in others much more limited. It is, then, the opportunity of our generation to seize the computer revolution, make it ours, and bring it forth for the good of all humanity.

Computational linguistics (CL) is a subfield of linguistics and computer science that is concerned with computer processing of human language. It includes automatic machine translation (MT) of one language into another, the analysis of written texts and spoken discourse, the use of language for communication between people and computers, computer modeling of linguistic theories, and the role of human language in artificial intelligence (AI).

The aim in MT is to feed into the computer a written passage in the source language (SL) (the input) and to receive a grammatical passage of equivalent meaning in the target language (TL) (the output). In the early days of MT, it was believed that this task could be accomplished by entering into the memory of a computer a dictionary of an SL and a dictionary with the corresponding morphemes and words of a TL. The translation-decoding program matches the morphemes of the input sentence with those of the TL. Unfortunately, what often happened
was a process called by early machine translators “language in, garbage out.” (Bharati, Chaitanya and Sangal, 1996: 108)

Translation is more than word-for-word replacement. Often, there is no equivalent word in the TL and the order of words may differ. There is also difficulty in translating idioms, metaphors, jargon and so on. Human translators deal with these problems because they know the grammars of the two languages in question and draw on general knowledge of the subject matter and the world to arrive at the intended meaning. MT is often impeded by lexical and syntactic ambiguities, structural disparities between the two languages, morphological complexities and other cross-linguistic differences.

According to Fromkin and Rodman (1995:473), the greater recognition of the role of syntax and the application of linguistic principles over the past forty years have made it possible to use computers to translate simple texts grammatically and accurately between well-studied languages. More complex texts require human intervention if the translation is to be grammatical and semantically faithful.

In general, a considerable number of problems that encounter CL and MT researchers may be traced back to the machine mistranslating of MT systems. These problems stem from the serious consequences of the incompatibility of the output with the input. Al-Asali (2000:xix) points out that the real problem with today’s MT systems, for example, ‘Power Translator’ and ‘Al-Wafi’, is that they do not achieve the appropriate interpretation of certain parts of the source text (ST), which may depend, in one way or another, on the appropriate comprehension of the devices controlling them.

Arnold (1994:268) indicates that the purchase of an MT system is in many cases a costly affair and requires careful consideration. It is important to understand the organizational consequences and to be aware of the system’s capacities. Unfortunately, it is not possible to draw up a comparison table for MT systems on the basis of which MT buyers could choose their system. Although system specifications can provide us with some useful information, there are too many aspects which influence the performance of MT that cannot be included in such a table. Furthermore, MT will perform differently in different translation environments, depending mainly on the character of the typical input texts. Consequently, if we want information about an MT system, we have to evaluate it, and that this evaluation has to be specific for the user’s translational needs which are the interest of the present study.
While there is a general agreement about the basic features of MT evaluation (MTE), there are no universally accepted and reliable methods and measures, and evaluation methodology has been the subject of much discussion in recent years (Hutchins, 2001a: 5-20). The main thrust of the work is to build up schemes that classify various aspects of import for MT, including user needs, the suggested system characteristics and associated metrics for measuring each of these. The classification schemes work on software evaluation.

As in other areas of natural language processing (NLP), three types of evaluation are recognized: a.) adequacy evaluation, i.e., to determine the fitness of MT systems within a specified operational context, b.) diagnostic evaluation, i.e., to identify limitations, errors and deficiencies which may be corrected or improved and c.) performance evaluation, i.e., to assess stages of system development or different technical implementations. Adequacy evaluation is typically performed by potential users and/or purchasers of systems, diagnostic evaluation is mainly the concern of researchers and developers, and performance evaluation may be undertaken by either researchers/ developers or by potential users (ibid.).

MTEs typically include features not present in evaluations of other NLP systems: the quality of raw (unedited) translations, intelligibility, accuracy, fidelity, appropriateness of style/register, the usability of facilities for creating and updating dictionaries, for post-editing texts, for controlling input language, for customizations of documents, etc; the extendibility to new language pairs and/or new subject domains, and cost-benefit comparisons with human translation performance (ibid.).

Initially, MTE was seen in terms of comparisons of unedited MT output quality and human translation. Later, systems were assessed for quality of output and usefulness in operational contexts. Subsequently, many potential purchasers have conducted their own comparative evaluations of systems, often unpublished, and often without the benefit of previous evaluations. Valuable contributions to MTE methodology which proposed evaluation tools for both system developers and potential users have been made. The initial intention to measure the “productivity” of systems for potential users is abandoned because it introduces too many variables. Evaluation, therefore, has concentrated on the performance of the core MT engines of systems, in comparison with human translations, using measures of adequacy (how well a text fragment conveys the information of the source), fluency (whether the output reads like good English, irrespective of accuracy) and comprehension or informativeness (using something like multiple choice tests covering the whole text) (ibid.).
Thus, with the rapid growth in sales of MT software and the increasing availability of MT services over networks, there is an urgent need for MT researchers, developers, and vendors to agree and implement objective, reliable, and publicly acceptable benchmarks, standards and evaluation metrics.

1.1.2 Evaluation of Existing Arabic MT Systems

It is only since about 1992 (ATA Software Technology, 1997: 7) that some Arabic MT systems have achieved certain degree of marked acceptance. There seems to be much confusion concerning the level of capability possessed by these products. At one end of the scale are mass media reports on these products that use the latest technique to produce near-perfect Arabic translations. Unfortunately, such reports are, generally, based entirely on the manufacturer’s promotional press releases without any attempt of actual evaluation and reliable verification of system’s quality.

At the other end of the spectrum are detractors of MT, those who steadfastly assert that such translation programs are useless, and the whole effort is a meaningless waste of time. Language translation is inherently too complex, they charge, and too dependent on human culture to ever be automated. This group, not surprisingly, includes a lot of working translators.

In the middle, however, is a much larger group of people who hold that MT technology, while not perfect, has progressed rapidly. Some of today’s systems, they believe, can render some English texts and documents into a very rough, but understandable, Arabic translation which can then be heavily post-edited by a translator.

In order to adequately answer the question of whether Arabic MT of texts and documents is practical, it is necessary to rigorously evaluate the currently existing Arabic MT systems at the local market. These are: Golden Al-Wafi 1.00 (2002); Al-Mutarjim Al Arabey 3.00 (2002) and An-Nakel Al-Arabi 2.00 (2000).

However, the evaluation of MT systems is a complex task. This is not only because many different factors are involved, but because measuring the overall translation performance of an MT system is itself difficult (Arnold, 1994: 270).
1.2 Problem

The present study raises the following questions regarding the evaluation of the available Arabic MT systems:

1. What are the types of evaluation that are to be applied to test the Arabic MT systems in the present study?
2. How Arabic systems are to be evaluated?
3. What are the translation tasks and user requirements these systems are supposed to achieve?
4. What system’s quality characteristics are to be selected for the present evaluation?
5. How well the selected quality characteristics are to be measured and evaluated?
6. What scoring schemes, metrics and methods are suitable for assessing system’s quality characteristics?
7. Is there any generally approved and well motivated scoring scheme?
8. What are the score ranges for ‘good’ and for ‘bad’ systems?
9. How are the evaluation results to be interpreted?
10. Which text type(s) does an Arabic MT system translate better and which does it translates worse?
11. Which Arabic MT system translates certain text types better than the other systems?
12. Which system will obtain the highest functional and/or non-functional performance?
13. Which system will obtain the highest total performance?
14. What are the merits and demerits of each of the Arabic MT systems to be evaluated?
15. Which system is better in practice? And how well it achieves the user’s needs?
16. What’s the effect of the user dictionaries on the translation quality of the system?
17. For what purpose(s) can the Arabic MT systems be used?
18. What types of problems do the Arabic MT systems encounter?
19. What sources of problems cause system’s deficiencies?
1.3 Hypotheses

It is hypothesized that:

1. Adequacy and performance evaluation types are the most suitable to investigate the Arabic MT systems’ capabilities and user requirements.

2. Evaluating the functional criteria of ‘suitability’, ‘accuracy’ and ‘wellformedness’ is the best way to obtain reliable results about the quality of the MT output.

3. Testing the non-functional criteria of ‘speed of translation’, ‘adaptability’, ‘storage’ and ‘flexibility’ is so essential to obtain more reliable results about the operational capabilities relevant to the translation process and systems’ quality.

4. The Arabic MT systems under evaluation are expected to perform a variety of tasks that fulfil the user’s requirements.

5. Depending on human judgement using scores on graded scales (e.g., 0-3 and 1-5) with different metrics and values of good and bad systems is the most effective method for the evaluation of quality characteristics.

6. The evaluation framework of the International Standards of Language Engineering (ISLE) (FEMTI, 2003) is the most effective, general and systematic model to be applicable to the evaluation of any MT system (whether Arabic or non-Arabic).

7. Discussing the evaluation results in terms of criteria types and text types where two evaluators are involved using tables, graphs and figures, in addition to statistical procedures, give reliable results of systems’ quality and performance.

8. As there are more dictionaries involved in the translation process (i.e., user dictionaries, preference dictionary, idioms dictionary, etc), it is assumed that the quality of translation produced by MT systems with such potentialities is superior to that resulted from systems that lack such capabilities.

9. Arabic MT systems are strictly aids that can support the translation process by taking over the parts of translation that are tedious and repetitious.

10. Arabic MT systems are not suitable for all text types and general translations as their manufacturers claim. In other words, the currently available general purpose MT systems cannot translate all texts reliably, i.e., output can sometimes be of very poor quality.

11. Favorite domains of application for MT can be made out in particular formally disciplined and highly standardized scientific and technical writing;

12. The major problems Arabic MT systems suffer from are the linguistic ones.
13. The main sources of problems with Arabic MT systems are due to their linguistic components.

1.4 Aims

The present study aims at:

1. Identifying and describing the various evaluation types and standards, and the stages of the evaluation process.

2. Working toward an exploration and application of a suitable methodology for evaluating three MT systems dealing with translation from English-into-Arabic, merely.

3. Selecting certain functional and non-functional criteria for the application of each measure and how the appropriate value can be determined for a given measure and a given system.

4. Developing a general framework in which existing evaluation measures for particular MT systems can be formulated in a systematic and organized way.

5. Discovering the extent of adequacy, as well as, the deficiencies of Arabic MT systems with respect to the quality of translation and systems’ performance.

6. Presenting diagnostic information about where a given system succeeds or needs improvement, relative to its intended users and use.

7. Providing comparative information which allows identifying the best system with respect to the translation quality and performance.

8. Specifying through the analysis of the results of evaluation the sources of problems that are responsible for producing ill-formed translations and inadequate systems’ performance.

9. Outlining some recommendations that are useful for system’s designers and developers to overcome various linguistic and operational problems that might impede the translation process.

1.5 Scope

The common practice of selecting just a few individual aspects to measure and evaluate has the danger of overlooking possibly relevant measures and of obtaining an inaccurate picture of the capabilities of a system. Since the types and number of MT evaluation criteria involving: a.)
general software characteristics, b.) characteristics of the translation process and c.) characteristics of the translation purpose, all with many subdivisions of further criteria, are too broad; the present study is confined to measure and evaluate only selective criteria of the general software characteristics that are most essential and relevant to the translation process, leaving the rest for further research.

Moreover, the study concerns itself with the evaluation of three uni-directional MT systems (i.e., that can perform the translation of various types of texts from English-into-Arabic, merely) using the available versions of MT software: Golden Al-Wafi 1.00 (2002); Al-Mutargim Al-Arabey 3.00 (2002); and An-Nakel Al-Arabi 2.00 (2000).

### 1.6 Procedure

The procedure to be followed in this study consists of two parts: theoretical and practical.

1. The theoretical procedures can be summarized as follows:
   a. Giving a full overview of MT, MTE, the characteristics of the translation process, general software characteristics, types of text and other relevant issues like reviewing the past evaluation practice and descriptions of various Non–Arabic and Arabic MT systems.
   b. Presenting a detailed identification and explanation of the Framework for the Evaluation of MT in ISLE (FEMTI, 2003) on the basis of which the current evaluation is based.
   c. Identifying and describing the properties of the available versions of MT systems that are capable of performing English/Arabic translations.
   d. Surveying some problems and issues pertaining to MT and evaluation of MT software.

2. The practical procedures, on the other hand, involve adopting a rigorous objective method of analysis and evaluation that concern:
   a. Establishing a general framework for evaluations of the participating Arabic MT systems involving classifying both the aspects of the system to be measured and the appropriate measures to use.
   b. Identifying a unified evaluation score in a systematic and logical way for how well the system achieved the user’s purpose.
   c. Classifying and applying comprehensive and accessible evaluation
methods for measuring system criteria for an adequate evaluation.

d. Achieving functional (i.e., cognitive and linguistic) and non-functional analyses and comparisons of various MT text types in terms of input/output with respect to the Arabic MT systems in question.

e. Testing MT quality using MT scale and MTE metrics.

f. Supporting the evaluation with statistical procedures, tables, graphs and percentages to establish an adequate MT quality assessment of each system and text type.

g. Identifying errors in MT with respect to each system.

h. Comparing Arabic systems’ performance and the characteristics regarding selected linguistic and non-linguistic criteria for the overall MT quality.

i. Discussing the merits and demerits of each system on the basis of the results obtained from the analyses, comparisons and evaluations. This involves isolating some significant system strengths, weaknesses, shortcomings and gaps.

j. Providing insights into the difficulties and subtleties of MTE.

k. Identifying on the basis of the above methods and evaluation procedures the best Arabic MT system which allows the user to make an informed choice of the suitable software.

l. Offering on the basis of the measurements, comparisons, and evaluations some suggestions that can help system builders and designers to find ways of system improvement.

### 1.7 Significance

With the increasing number of MT systems available on the market, the demand for valid ways of assessing their translation capabilities from the users’, as well as from the developers’ point of view becomes even greater. Evaluation is one of the most important stages in the life of any software. This is not only for its potential users and buyers, but also for researchers and developers. Evaluation is also crucially important to scrutinize any MT system, analyzing the quality of its output, classifying errors it makes and improving it. The evaluation can, thus, be
concerned with the technical quality of the system. It can also target the computational and linguistic limitations of the system, the software engineering aspects and / or the cost and benefits.

In fact, MT users need to know which system is appropriate for their specific requirements. This cannot, of course, be decided by the layman by simply saying this or that system is good or bad. Evaluators, who are, in most cases, translators, system designers or researchers can perform various tests based on objective and scientific evaluation standards to indicate the quality of such systems. So, the present research is significant for it concerns itself, through the application of special metrics and methods, with the evaluation of different Arabic MT systems available, taking into account a variety of quality criteria and text types. Thence, it points out to: a.) whether an MT system is good?, b.) which of two or three systems is better for a certain text types? and c.) What do ‘good’ and ‘better’ mean in this context?

Thus, a study like the present one is of value for people who want to compare several MT systems. They can browse and select the characteristics that best reflect their circumstances by finding associated evaluation measures and tests. It is also useful for people who want to design a new MT system or upgrade an old one as they can learn about the needs of users and find niche applications for their system.

Above all, to the best of the researcher’s knowledge, this study is the first of its kind in detecting the evaluation of and comparing recently released new versions and new MT systems dealing with English-into-Arabic within the framework of the formal international standards of evaluation. This is to:

1. make sure to what extent the claims of the manufacturers of these systems are true as they are meant for advanced and professional levels to translate a variety of scientific and humanity texts,
2. investigate the strengths and weaknesses of them and types of errors for subsequent system’s improvements by designers and developers, and for designers of MT systems in general to avoid the problems of the present ones,
3. indicate the economic value of such systems as they are very expensive with respect to the quality of translation they perform,
4. be in line with the international advancements and research projects in the field of MT and MTE and
5. determine where and how they could be utilized by the end-user.
This whole question of evaluation of MT, in fact, has always been a hot topic especially in recent years, and it is hoped that the present study will be a good contribution in this field.

1.8 Plan

The present study consists of six chapters and a list of references:

The present study consists of six chapters and a list of references:

The first chapter (Chap. 1) is devoted to the identification and scope of the problem of evaluating Arabic MT systems. Moreover, it outlines the hypotheses, the aims of the study, the theoretical and practical procedures of the study. It also indicates the value of the study for designers and developers for system’s improvement, as well as for those who are concerned with using these systems to perform various translations and relevant tasks.

Chapter two attempts to focus on the scope and role of evaluation in MT, presenting some past and recent evaluation studies, standard work related to evaluation and new trends and future work in MTE. The chapter also introduces types and purposes of evaluation, and the state of Arabic MT and MTE. In addition, it gives description of Arabic MT systems, in general focusing more on the systems evaluated in this study.

The third chapter is devoted to the identification and discussion of: a.) the level of evaluation, b.) the model of evaluation and c.) the various stages of the evaluation process. The aim is to produce a framework on which the analyses of the Arabic MT output and performance are based.

Chapter four deals with the description of the experimental work. It is mainly concerned with:

a.) The experimental design involving text sampling, the systems tested, the evaluators, user and task description.

b.) A user –oriented model of test types regarding systematic testing and testing instruments.

c.) The evaluation procedure which discusses the implementation of evaluation methodology of testing the selected functional and non-functional criteria.

The fifth chapter includes the classification and discussion of the evaluation results obtained by testing the three Arabic MT systems in terms of various criteria types and text types (i.e., twelve texts representing 268 sentences for each MT system) to find out the overall quality for these systems. A detailed discussion of these findings supported with illustrative examples
culled from the data is given. The discussion covers the specific reasons of the cognitive, linguistic and operational problems of these MT systems.

Chapter six presents the final conclusions of the study based on the results and discussion offered in Chapter Five. Probable sources of problems are identified and discussed, and then some recommendations are offered. The chapter ends with few suggestions for further development and expansion of the present study.
CHAPTER TWO
EVALUATING MT SYSTEMS:
SOME BASIC ISSUES

Ever since the idea of using computers to translate natural languages was first proposed in the 1940s and since the first investigations were begun in the 1950s, translators have watched developments either in scorn, dismissing
the very notion that anyone could ever believe that translation could be mechanized, or in trepidation, fearing that their profession would be taken over entirely by machine.

However, there is now no doubt that computer-based translation systems are not rivals to human translators, but they are aids to enable them to increase productivity in technical translation or they provide means of translating material which no human translator has ever attempted.

In fact, more has been written about evaluating MT systems than about MT itself (Hovy, 1999). Yet, after 50 years there is still no standard test or test set, no internationally accepted rating system, and little agreement about what exactly should be measured.

Recently in the USA, the American Research Programme Association’s (ARPA) MT programme has included an annual evaluation contest, in which research systems, commercial systems and professional translators were pitted against one another (White et al., 1994). Not only was this exercise phenomenally expensive (the last one, involving 18 entrants, took over three months and cost over $400,000), the results were not very informative, neither to potential system users such as commercial or Government office; nor to researchers and system builders. Hovy (1999) indicates that “ideally, MT systems and results should be simple to evaluate without requiring grossly oversimplified measures.”

It is worth mentioning that evaluation of language tools, particularly tools that generate language, remains an interesting and general problem. MT is a prime example. Approaches to evaluating MT are even more plentiful than approaches to MT itself; and the number of evaluations and range of variants is confusing to anyone considering an evaluation. Unfortunately, many prior evaluations do not include an adequate
specification of important aspects such as evaluation process complexity, cost, variance of score, etc. (NAACL, 2001).

In an effort to systematize MTE and drive it to the next step, many projects such as the Expert Advisory Group on Language Engineering Standards EAGLES Report of 1996; the EAGLES Report of 1999; and the ISLE Project 1999-2003 have created taxonomies of evaluation-related features and measures with associated pointers to systems.

The present chapter focuses on the importance of MT systems and their evaluation, types of MTE, previous and very recent MTE research, types of MT systems, description of MT systems under investigation, the principles of classification scheme for MTE, standards for software evaluation in addition to some difficulties of MTE and other relevant issues of evaluation.

Finally, it is important to know how to measure progress during the development of an MT system. It is also important for users to be able to measure which system or which combination of systems best suits their needs.

2.1 General Overview

MTE has been more “magic” and “opinion” than science. The history of MTE is long and checkered and the search for objective, measurable, resource-reduced methods of evaluation continues (Vanni & Reeder, 2001). In this respect, Gross (1999:3) confirms that:

Methods for evaluating machine translations and machine translation systems have been proposed, discussed, and applied for more than 40 years now, including numerous attempts at defining objectively measurable criteria to capture aspects of translation quality. Nevertheless, a worryingly large number of evaluation reports have
more or less explicit disclaimers as to the absolute value of the results, or confess to flaws in the procedure.

The obvious solution to these problems is, of course, to avoid translation quality as a direct object for evaluations and to stay with a general impression of the role which quality plays for the overall acceptability of an MT system.

Moreover, Hutchins (1997:418) maintains that while there is general agreement about the basic features of MTE as reflected in general introductory texts (Lehrberger & Bourbeau 1988, Hutchins & Somers 1992, and Arnold, 1994); there are no universally accepted and reliable methods and measures, and evaluation methodology has been the subject of much discussion in recent years (See also AMTA, 1992 and Falkedal, 1994).

According to Vanni and Reeder (2001), the notion of evaluating MT products results is too broad of a scope for reasonable evaluation–everything from interface, to scalability, to faithfulness of translation; to mean–time–between–failures of the system is fair “game” for the evaluation of MT systems. Yet, it is necessary to have a method to measure the usefulness of a system to users and equally desirable to point to places where system designers and researches can improve system outcomes.

Ideally, it is hoped to improve the performance of MT systems by improving the system, but it might be even more important to improve performance by looking for a more appropriate application. A survey of the literature on evaluation of MT systems seems to suggest that the success of the evaluation often depends very strongly on the selection of an appropriate application. If the application is well-chosen, then it often becomes fairly clear how the system should be evaluated. Moreover, the evaluation is likely
to make the system looks good. Conversely, if the application is not clearly identified (or worse, if the application is poorly chosen), then it is often very difficult to find a satisfying evaluation paradigm (Church & Hovy, 1993: 239).

As is proved by the increasing availability of commercial MT and machine-aided translation (MAT) systems, MT today is beginning to find areas of real commercial applicability. Thus, to the question “How can one build MT systems that make a difference?” the answer is that the community needs to find evaluation measures and applications that highlight the value of MT research in those areas where systems can be employed in a real and economically measurable way (ibid.). In order to design and build a theoretically and practically productive MAT system (as human and machine translation show complementary strengths), one must choose an application that exploits the strengths of the machine and does not compete with the strengths of the human. This point is well put in the following:

The question now is not whether MT is feasible, but in what domains it is most likely to be effective… The objective of an evaluation is, of course, to determine whether a system permits an adequate response to given needs and constraints.

(Lehrberger & Bourbeau, 1988: 192)

In spite of all the literature on MTE, the general evaluation measures that are proposed often fail to point the strengths of systems and lead them toward real utility; instead, they seem to confound important and less important aspects.

Evaluating MT is important and is “one of the most important issues in this field” (Yokoyama, 2001). Researchers need to know if their theories
make a difference, commercial developers want to impress customers, and users have to decide which system to employ. As a result, the literature is replete with MTE and evaluation studies. However, just as it is nonsensical to ask “what is the best house?” it is nonsensical to ask “what is the best MT system?” No simple answer can be expected or given when evaluating MT systems, even poor MT can be useful or even ideal, in the right circumstances (Hovy & King, 2001: 1).

The overall process of evaluation is the same whether comparing different systems or trying to evaluate a single candidate system. The ultimate question is whether it fits with what the customer of the evaluation wants or needs.

To sum up, evaluation has always been central to the consciousness of those involved in the field of MT. Historically, evaluation has proven difficult, traumatic, at times misleading, but very often both revelatory and helpful. Originally, it was the apparent results of evaluations themselves which made the general public aware of the potential for MT. Today, there is emerging a legacy of actual production use of the output of MT from which the fuller understanding of its potential becomes apparent to actual users. The increased usage of MT, however, demands more comparability and relevance among the many attributes and measures of MT (White in Somers, in press).

### 2.2 Scope of Evaluation

Generally speaking, automatic translation has not yet achieved a breakthrough, and apart from a very few users it is still at the research and development stage. Hence, it is important to make an evaluation of its
performance and potential to determine whether development work on these systems will be continued or abandoned.

Evaluation is the measure of the quality of a system, in a given context, as stated by the definition of quality as “the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs” (ISO/IEC-9126, 1991: 2). It is the process of system appraisal which leads to global, overall, quantification of performance. It is concerned with detailed measurements involving the analytical description of system performance in terms of defined factors (Assessment and Evaluation Defined, 2002). Also, it comprises validation (test against implied needs, i.e., assumptions), verification (test against stated needs, i.e., specification), measurement (mapping of an attribute onto real numbers) and assessment (comprising actual measurement results again required) (Hausen: 2002).

Evaluation projects and campaigns aim at defining techniques and standards measuring usefulness, success and efficiency of software component with respect to MT between two or more human languages. The quality of a NLP system is estimated in terms of: a.) a given task, b.) a given category of users and c.) economic factors. Revising ghosts of the past, the reasons of MTE failure may not apply now as corpora availability and processing power has increased and expectations have changed (What is MT Evaluation: 2002).

Hovy (1999) points out that “what the user needs [is] the ability to characterize his or her particular needs (personal and organizational), and the ability to compare this characterization with the performance characteristics of various MT engines”. He (in An Evaluation: What to Do and What not to Do, 2001) adds that the best thing an evaluator can do is to describe his/her
needs, environment and give some sample text; while the worst thing an evaluator can do is to apply some arbitrary measure without understanding what it measures, without knowing when it is applicable and without understanding the accuracy or reliability of the metric. What is needed is a systematic outline of the metrics and their appropriateness and their scope. His conclusion indicates that “MTE is pain, expensive, stressful and inconclusive. What’s needed: a project of evaluating evaluations.”

King (1996: 73) maintains that

I find it very hard to talk into empty space about what counts as a good or a bad translation. I need to know what it’s for and what the criteria are in that particular situation before I can even talk about evaluating a translation.

When comparing a number of existing MT systems, it is important to know how wide the bounds of the evaluation should be set and what the scope of the requirements is. When a software system is placed within a human process, as with many interactive systems, it may be relevant to evaluate the human plus software system as a whole, or at least to aim at validating and ranking requirements of software quality in terms of their correlations with the performance of the system as a whole. For instance, in the case of spelling checkers, it must be relevant to the impact of spelling checkers on the overall task of document quality assurance to know that the errors that spelling checkers are able to spot are also those that people find it easiest to spot (EAGLES Report, 1996: 25).

Traditionally, evaluations have involved qualitative measures, such as ‘faithfulness’ to the original message; ‘intelligibility’ and
‘comprehensibility’; accurate rendition of terminology; and stylistic appropriateness (for the specific language and subject). Increasing use is made of measures for evaluating utility or usability: the savings of costs and time, in comparison with other systems or with wholly human translation; the ease of use; the level of intelligibility and/or accuracy, in accordance with the intended or expected application; speed and response times; training and setup costs; impact on an organization’s overall translation throughput; and compatibility with other systems. It has always to be stressed that systems suitable and cost-effective for one particular organization or individual may be quite unsuitable and uneconomic for another organization or individual (Hutchins, 2000: 10).

Finally, evaluation involves much more than the quality of translation (although that is obviously a most important aspect). It involves also, for example, the integration of an MT system in the whole processing framework: transmission and receipt of texts, formatting, dictionary updating, editing, printing and distribution. It involves examination of the compatibility of systems with other computer facilities, and in particular it embraces the integration of the system into the working patterns, practices and attitudes of existing staff, and the aims of the organization as well.

2.3 The Role of Evaluation in MT

The evaluation of MT systems has been a central research topic in recent years (Sparck-Jones and Galliers, 1995 and King, 1996). According to White in Somers (in press), there are three reasons for the primacy of evaluation in the MT field. The first indicates that MTE is central today, because it was central once and it is just focused on today, because it has always been so. It
is of great value to regard the issues raised in the earliest days of the field as remaining critical today.

The second reason why MTE is pre-eminent is because it remains important today. It costs a great deal of money to research, design and implement an MT system, and more time and money still to complete the system with “knowledge” (e.g., words, phrases, meanings, contexts, etc.) germane to the subject areas which the system will translate. The different interests in MT need to know whether the investment is worth making for their individual objectives (ibid.).

That MTE is hard to do, constitutes the third reason. Translation is special among the set of automated applications that may be called “human language technologies”, because “correct translation” is an elusive target, and because there are a range of people, purposes, and types of MT that each needs different measures to indicate what each needs to know about MT systems (ibid.).

It is axiomatic that evaluation measures some attribute of something against a standard for that attribute. For this to happen, there needs to be an identifiable “correct” or “best” ideal, whether explicit or implicit, against which to compare the relevant attribute of the individual item being measured. The most obvious standard for MT, i.e., the “right” translation is the very thing translation itself cannot provide. Actually, there are many ways to translate the same thing, and researchable translators will disagree about which way is best. These facts are a testimony to the rich variability of language and remarkable creativity that goes into the act of translating. But, it certainly makes life harder when trying to evaluate MT systems (ibid.). In the present study, the evaluation of some attributes requires comparison with an adequate human translation that acts as a standard for that attribute.
In short, MTE is hard because there is still no ‘gold standard’. Moreover, there is a wide range of parameters on which to evaluate, but not all have some importance to every user and the question remains “what is acceptable to a user?” Further, there is a wide range of uses of product and not all are as tolerant of failure. And, “MTE is hard, because MT is hard” (Why MTE is Hard? 2002).

2.4 Some History

2.4.1 Background

MTE has a long history. Even very early in that history, attempts were made to produce well designed and well founded evaluation schemes. Evaluating MT is important for everyone involved. Given the richness of the literature and the complexity of the enterprise, there is a need for an overall perspective, something that helps the potential evaluator approach the problem in a more informed way that might help pave the way toward an eventual theory of MTE.

MTE is a long-standing issue with many approaches and formalisms having been proposed throughout the years. What to evaluate, how to evaluate and what context to use in evaluation are problematic issues. Unlike some other NLP problems, there is no ‘gold standard’ evaluation possible. This lack of “ground truth” makes the task of automating evaluation even more challenging. The lack of agreement on the assessment of what makes a good translation, even when human translators are involved, hampered initial efforts in MTE, which compared the output of systems to renderings produced by professional translators (e.g. ALPAC, 1966; White et al. 1992-1994 (D)ARPA Workshops on MT). The results of tests of adequacy, informativeness, and fluency as performed on system output were compared
to the results of those performed on the human renderings. While this notion of focusing on the outcome of the translation process is a reasonable one, the implementation of the tests proved difficult and somewhat detrimental to the field. In order to assemble the amount of data necessary, such MTE programs were expensive, time-consuming and human-intensive (Vanni & Reeder, 2001).

Moreover, these programs (MTE programmes) measured only one broad aspect of the translation output at a time. Therefore, developers were left with little help to improve their systems and users were left with little guidelines to select an appropriate system to meet their requirements. For example, one finding of the Defense Advanced Research Projects Agency (DARPA) evaluation (DARPA, 1994) was that larger knowledge sources were correlated with better performance (White et al., 1994) – a useful piece of information in a general sense, but not particularly helpful for specific system designers or users. Further, in the past the evaluation of MT systems has focused on single system evaluations because there were only few systems available. But now, there are several commercial systems for the same language pair. This requires new methods of comparative evaluation (Vanni & Reeder, 2001 and Volk, 2001).

However, even though much that is contentious surrounds the work of the Automatic Language Processing Advisory Committee (ALPAC) in the mid 60s, there is no denying that the ALPAC Report (Pierce et al., 1966) contains a laudable attempt at good experimental design. Later, the Van Slype Report for the European Commission (Van Slype, 1979b) provided a very thorough critical survey of evaluations done to that data. Much valuable material can also be found in the Association for MT in the Americas AMTA (1992), in the OVUM Report of Mason and Rinsche (1995) and in the
accounts of the *(D)ARPA* MT Program (1992-1994), which contained a substantial evaluation component.

Popescu-Belis, Manzi and King (2001) state that one way to get a handle on the question of how an evaluation for an MT system should be designed, might be to take seriously the notion that judgement of a system critically depends on the ‘context’ in which use of that system is envisaged. This was the strategy adopted by the Japan Electronic Industry Development Association, *JEIDA Report* (Nomura and Isahara, 1992), which attempted to systematize the description of users’ present situation and their real needs.

It is worth noticing that plans for large-scale evaluations have become more functionally oriented. For example, the MT scale plan sought to associate the diagnostic scores assigned to the output used in the DARPA evaluation with a scale of language-dependent tasks such as scanning, storing and topic identification. Linking the breakdown in a user’s language-based performance of a function to some phenomenon in system output extended the usefulness of this approach. Similar types of associations were explored even further with experiments in correlation systems’ handling of a set of text feature, with user’s performance on information processing tasks and measuring a system’s performance on new text types. Actually, consideration of variables such as the function of MT output and the complexity of MT input continued to be explored by researchers with the recognition and description of the role of the user’s purpose and process (Hovy, 1994 and White & Taylor, 1998).

So far, this section has addressed MTE research and why it has been a difficult challenge. The next section gives a consideration of some past evaluations. No attempt is made at producing an exhaustive review. For a detailed critical review of a number of MTEs and NLP in general, both
Galliers and Jones (1993) include many valuable insights. The intention behind the choice of example evaluation, here, is to illustrate the variety of evaluation scenarios.

2.4.2 Some Past Evaluation

2.4.2.1 ALPAC

The *ALPAC Project* (1966) of MT was one of the first evaluations. The evaluation as a whole can be thought of as an adequacy evaluation (i.e., to determine what the potential customer’s needs really are), comparing MT to human translation on the three dimensions of speed, cost and quality. The measure used was to take a set of translations, some produced by machine, some by human translators, and ask a group of test persons to rate the translations on two scales, one for ‘intelligibility’ and one for ‘fidelity’. The use of rating scales subsequently became widespread in evaluation of MT (*EAGLES*, 1996:50).

The method was carefully designed. The test material was (144) sentences randomly selected from four different passages of a Russian book. Six different translations were produced for the (144) sentences (three by human translators and three by different MT systems). The translations were then merged randomly into six sets, with the constraint that each sentence appeared only in one translation in each set. Each set was then given to three monolingual and three bilingual test persons, all of whom had one hour’s training using a set of thirty sentences drawn from the same material as the test set. There were thirty-six test persons in total. A definition in English was given for each of the points on each of the rating scales. Thus, ‘intelligibility’ was rated on a nine-point scale from “perfectly clear and
intelligible” to “hopelessly unintelligible”. ‘Fidelity’ was defined over a ten-point scale in terms of informativeness (ALPAC, 1966, Appendix 10, 67-69). The committee reached extremely negative conclusions about what could be hoped for from MT systems in the short to medium term. Nevertheless, the ALPAC evaluation must be considered a pioneering effort if only because it emphasized the importance of good evaluation methodologies.

In general, for commercial NLP systems, it is rare to find a product which will do all and only what the customer wants. Frequently, the system will have to be modified or extended to meet specific needs. Thus, evaluation is aimed at finding out not only what the system currently does, but also how easily it can be modified (EAGLES, 1996:51).

2.4.2.2 (D) ARPA

From (1992) onwards the (D) ARPA sponsored a series of evaluations of MT systems. The report here is based on White et al., (1994). In the (ARPA, 1993) case, the needs were those of the funding agency. The declared aim of the research programme was to “further the core technology”. The funding agency therefore needed comparative evaluation of systems based on different technologies and translating from different languages into English. The difficulty of the task was further compounded by an invitation to operational systems (commercial or otherwise) from outside the research program to participate in the evaluation exercise. Furthermore, there were great differences in the way the systems were intended to be used. At one extreme, one system was planned as a fully-automatic batch-oriented
system, at the other was a system intended more as an on-line aid to a human translator than as translation system (EAGLES, 1996:51).

Given all these constraints, the only quality characteristic which offers any hope of comparability is functionality, and that only if it is interpreted in the widest sense to allow the output of a machine-aided human translation (MAHT) to be compared with the output of a fully automatic MT. Two attributes of functionality were picked out: a.) comprehensibility of the output, and b.) quality of the output. The test materials in the 1992 evaluation were constructed by taking a set of English newspaper articles about financial mergers and acquisitions, and having them professionally translated into the source languages of the different systems worked from. The systems then translated the translations back into English and it was these outputs that were evaluated (ibid.). The three subsequent ARPA evaluations have retained the comprehensibility attribute but have replaced the quality attribute with two sub-attributes, i.e., adequacy and fluency.

Reeder et al., (2001) point out that the 1994 ARPA MTEs tackled the issues of: eliciting dimensions of judgements from otherwise disinterested target-native subjects; capturing judgements with finer granularity than before; and using sufficient human factors controls to show reasonably valid measures of ‘fidelity and intelligibility’. In this series of evaluations, system outputs were compared to human translations on the criteria of adequacy, informativeness and fluency (the first two are measures of ‘fidelity’ while the last is of ‘intelligibility’). The criteria were elicited from subjects whose rating was on a holistic 1-5 scale. The ARPA evaluation was expensive and time-consuming because of the need for: a.) two expert translations of input texts and b.) an elaborate test
design and administration procedure, with a large number of input documents, output documents, human subjects and decision points for measurement and analysis. Moreover, the ARPA series did not directly provide insight into the place of MT in the continuum of language processing and NLP users (ibid.).

2.4.2.3 The Hewlett Packard Laboratories Study

Data base query is another application of NLP with a long history of evaluation. Woods (1973:441-50) described informal field testing of the LUNAR system through monitoring the treatment of (110) queries during demonstration of the system and Damerau (1980) reported more extensive field testing (a transformational grammar based) over a period of two years from late (1977) through (1979). Both of these were clearly adequacy evaluations, with the interesting characteristic of being executed in close collaboration with the end-user community. The emphasis on field testing of data base query systems was also reflected in two other works by Jark et al., (1985:97-113) and Whittaker and Walker (1989).

The main quality characteristic considered relevant for evaluation of a generic system (i.e., a system not specific tailored for use with one particular data base) was the functionality of the system. The relevant attributes were linguistic and computational. A “test suite” was constructed to provide data for various measures relevant to these attributes. The test suite consisted of a large number of English sentences annotated by a construction type. The sentences covered a wide range of syntactic and semantic phenomena, including anaphora and inter-sentential dependencies. Ungrammatical examples were included. Vocabulary was limited. The method was only described in very general terms: the sentences were processed by the system
being evaluated, the data base query was generated and the resulting query was used to query the data base. The results provided data relevant to a number of different measures. Most of these measures were intimately related to the theory incorporated into the system. Accuracy of parsing, for instance, could only be measured against what the theory of parsing implemented defined as a correct parse. The close connection between measures and theories underlying the system was typical of diagnostic evaluation (*EAGLES*, 1996:53-54). The purpose of the exercise was to provide feedback for the research workers developing the system on where modification or extension was needed.

2.4.2.4 Message Understanding Conferences (MUC-3)

In (1991) four conferences were held: Chinchar, Lehnert and Sundheim, Message Understanding Conferences (MUC-3) and Sundheim. The description here will concentrate on the third. The type of evaluation was black-box and the basic strategy was to define a goal by creating a test collection consisting of a set of text and a set of relevance criteria for the texts and for the information to be extracted. A set of “answer templates” was then defined, and the system was evaluated by comparing the templates it produced with the answer templates. Fifteen systems participated in MUC-3. The evaluation was thus a comparative evaluation of these systems’ adequacy in fulfilling the particular task. Re-testing a system after modification using the same material could also produce an evaluation of that system’s progress (ibid: 54-55).
The only quality characteristic taken to be relevant was the system’s functionality. The attribute was the system’s ability to extract essential information of the specified kind and the measure was the number of template slots correctly filled in each case. MUC-3 used a corpus of 1,600 articles each about half a page long drawn from a variety of text types. The articles were divided into a training set of 1,300 texts made available to all fifteen participating cite and a test set of 300 articles. The articles covered a wide range of linguistic phenomena and included ungrammatical input. A semi-automated scoring program was developed to calculate the various measures of performance. The two primary measures were ‘completeness’ and ‘accuracy’. ‘Completeness’ was calculated as the ratio between the number of template slots filled correctly by the system and the total number of filled slots in the answer template. Fills corresponding exactly to the fill in the answer template scored 1.0, whilst fills judged by humans to be a good partial match scored 0.5. ‘Accuracy’ was the ratio of slots correctly filled to the number of total fills generated. Two other measures, ‘over generation’ and ‘fallout’ were also used (ibid.).

A look at the results shows that, in general, all systems perform better on ‘accuracy’ than on ‘completeness’. Eight sites achieved at least 20% ‘completeness’ and 50% ‘accuracy’. Two systems exhibited ‘completeness’ scores over 40% with ‘accuracy’ over 60% (ibid: 56).

Although comparative evaluation provides clear information on a system’s adequacy with respect to the particular task in hand, it is difficult to estimate the usefulness in predicting the capacity of individual systems to evolve and progress or the portability of systems trained on texts drawn from one domain to a different domain (ibid.). It is worth noticing that applying ‘black-box’ evaluation comparatively to a number of different systems
working on different principles is now also being adopted with ongoing research on MT.

In short, evaluations vary enormously in their purpose, in their scope and in the nature of the object being evaluated. Consequently, it is hardly surprising that evaluation techniques in their turn differ widely, as do the resources they require. In addition, it is in principle impossible to envisage the design and construction of some general evaluation tool, into which any NLP system could be plugged in order to obtain data relevant to a set of informative measures.

2.4.2.5 COBALT

COBALT is another evaluation research project, collaborated with the TEMA project in an attempt to evaluate the EAGLES/TEMAA framework proposals by confronting them with the evaluation methods actually used in the COBALT Project. News articles in the financial domain were automatically analyzed and classified at two levels of classification to discover whether they were of potential interest to either one of two user communities. The evaluation methods used for the different components of the system and for the system as a whole were analyzed using the EAGLES framework as a way of structuring the analysis. The exercise proved very useful to all parties and confirmed that the EAGLES way of thinking about evaluation provided a constructive approach to designing evaluation methodologies also in the context of project evaluation (EAGLES, 1996: 48).

To sum up, the path to a systematic picture of MTE is long and hard. While it is impossible to write a comprehensive overview of the MTE literature, certain tendencies and trends should be mentioned. Throughout the history of evaluation two aspects stand out, i.e., quality and fidelity.
Particularly, MT researchers often feel that if a system produces syntactically and lexically well-formed sentences (i.e., high quality output), and does not distort the meaning (semantics) of the input (i.e., high fidelity) then the evaluation is sufficient (Hovy & King, 2001: 3).

### 2.5 Standards Work Related to Evaluation

The International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) together form the specialized system for worldwide standardization. ISO and IEC Committees collaborate in fields of mutual interest. An important standard pertaining to evaluation is ISO/IEC 9126 (1991). The first edition of this standard, entitled *Information Technology-Software Product Evaluation-Quality Characteristics and Guidelines for their Use* was published in 1991. This standard was mainly concerned with stipulating a set of quality characteristics for software. Similarly, a new standard ISO/IEC 14598-1, which gives a general overview of the process of evaluation, has only recently been published as an international standard (*EAGLES*, 1999: 5-6).

It is worth noticing that the *ISO/IEC 9126* (1991) distinguishes between internal characteristics which pertain to the internal workings and structure of the software and external characteristics which are the characteristics which can be observed when the system is in operation. MT systems are taken as a practical example of the ISO-based framework for evaluation, but the methodology should apply to any software application falling within the domain of human language technology at any stage of its life cycle, from initial conception to commercialized product. Indeed, the general principles of the framework may apply to any complex object. Recently, a classification has been developed which brings together a wide range of
evaluation criteria, i.e., the ISLE (2000) classification framework for evaluation of MT. It provides classification of the internal and the external characteristics of MT systems to be evaluated in conformity with the ISO/IEC 9126 standard concerning quality characteristics of software products. The ISLE project derives a lot of its work and vision from the EAGLES I (1996) and EAGLES II (1999) Projects. The EAGLES guidelines for evaluation are themselves inspired by the ISO 9126 standard for software evaluation, with links and references to the further ISO 14598-1 (1998) standards (Marrafa and Ribeiro, 2001) and (Popescu-Belis, Manzi and King, 2001).

The ISLE Project-Evaluation Work Group organized series of workshops for a practical work using the ISLE taxonomy (the full version available over the Internet (http://www.issco.unige.ch/projects/isle/taxonomy2, 2001). There has been considerable continuity between workshops, with the result that the most recent in the series offered a number of interesting examples of using the taxonomy in practice. A very wide range of topics was covered, including the development of new metrics, investigations into possible correlation between metrics, ways to take into account different user needs, novel scenarios both for the evaluation and for the ultimate use of an MT system and ways to automate MT evaluation. The four workshops took place in October 2000 (at AMTA 2000); April 2001 (at ISSIDO, Geneva); June 2001 (at NAACL 2001) and September 2001 (at MT Summit VIII, Spain).

In a very recent workshop at the Language Requirements Evaluation Conference (LREC) 2002 in Canary Islands- Spain, Hovy, King and Popescu Belis (2002) describe the principles and mechanism of an integrative effort in MTE. Building upon previous standardization
initiatives, above all ISO/IEC 9126, 14598 and EAGLES, they attempt to classify into a coherent taxonomy most of the characteristics, attributes and metrics that have been proposed for MTE. The main articulation of this flexible framework is the link between a taxonomy that helps evaluations define a context of use for the evaluated software, and a taxonomy of the quality characteristic and associated metrics. The document presented by them overviews these elements and provides a perspective on ongoing work in MTE.

In subsequent sections and subsections of the present chapter, previous approaches to MTE will be outlined with focus on the description and classification of the taxonomy developed by Hovy, King and Popescu-Belis (2003), since it is the most recent work based on and supported by the ISLE Project. Moreover, a brief overview of the EAGLES 7-major steps necessary to carry out a successful evaluation of language technology systems or components is also presented (these 7-steps are introduced in EAGLES, 1999).

2.5.1 Previous Approaches to MTE

Much work has been done on evaluation of MT in the last ten years. A common goal has been the design of evaluation techniques in order to reach a more objective evaluation of MT quality systems.

In this section, only certain tendencies and approaches to MTE are discussed, since it is impossible to give a detailed account of all the standards and procedures of MTE.

2.5.1.1 JEIDA (1992)
The Japanese JEIDA study (Nomura and Isahara, 1992:1-17), paralleling EAGLES observed that often the user’s present situation and the user’s true needs differ. Two questionnaires were created, situation and needs, each focusing on 14 parameters. The user’s answers to the questionnaires were plotted on two radar plots. Superposition of two plots provided an immediate and striking visual answer to the overlaps and differences for each aspect. Using these differences, then, the user could determine which of the seven MT systems (also characterized on radar plots) would be best suited to his situation (Hovy, 1999 and Koh et al., 2001).

The JEIDA showed a neat method that relates user’s needs to MT systems. Their approach allows each user to determine the relative importance of the various aspects of MT for himself. However, the method is limited by which features the JEIDA evaluators decided to include into the questionnaires and radar plots. This can be a problem, as the authors admit:

This set of criteria permits both the overall evaluation of the system and the evaluation of technical components which have been incorporated into the system. As a result, some advantages or disadvantages of the system, either general or specific, can be recognized by the developer. However, this set of criteria does not evaluate the quality of MT systems directly.

(Nomura and Isahara, 1992: 14-15)

This problem-limited feature set is endemic to all MTE schemes. In some aspects, the features may provide too much detail for the user’s current needs; in other aspects, too little. One may ask: what’s wrong with getting all the data? Indeed, too much information can be as harmful as too little. For instance, the *OVUM Report* (1995) attempts to provide a comprehensive
characterization of MT systems where each individual system is discussed according to usability, customizability, use within total translation process, language coverage, processing technology, ease of adding terminology, completeness of linguistic resources and utilities, documentation, user support and training, etc. To a non-sophisticated user, this profusion may be confusing (Hovy, 1999).

2.5.1.2 The EAGLES Evaluation Guidelines

The European EAGLES initiatives (1993-1996) came into being as an attempt to create standards for language engineering. It was accepted that no single evaluation scheme could be developed for a specific application, simply because what counted as a “good” system would depend critically on the use of the system. The idea that user requirements can be specified and used as a tool in designing an evaluation also lies behind the EAGLES work on evaluation. The first EAGLES Project (1996) was set up to meet this need (EAGLES, 1999:6).

Influenced by the earlier work on evaluation, a primary insight during the first EAGLES period was that it was impossible to come up with one general recipe that would be valid for all evaluations in all circumstances. It was however believed that it should be possible to create a general framework for evaluation design, which could guide the creation of individual evaluations and make it easier to understand and compare the results. It was quickly realized that adopting this point of view also fitted well with work done by ISO/IEC 9126 in 1991 on creating standards for the evaluation of software (Hovy and King, 2001: 2 and Popescu-Belis, Manzi and King, 2001).
Potentially, the pre-eminence of a user with specific needs is in tension with the desire to build an evaluation framework of general validity. The EAGLES group resolved this tension by thinking in terms of classes of users, groups of individuals who could be considered to share certain needs. This leads to the notion that in any given context, not all quality characteristics are of equal importance. In other words, it may well be that in one context, reliability is critical and far more important than speed, while in another, speed in obtaining even imperfect results outweighs both accuracy and reliability (Popescu-Belis, Manzi and King, ibid.).

By pooling the requirements of each of the classes of users, it should be possible to come up with a general quality model. Such a model will be a structured collection of quality characteristics, usually broken down into sub-characteristics, bottoming out in a set of measurable attributes. Each measurable attribute is accompanied by one or more valid metrics, which yield a score for that attribute when the evaluation is executed. The scores thus achieved can be combined in ways that reflect the relative importance of the attributes in a specific evaluation. Indeed, when designing a specific evaluation, it may be decided that score attributes are of no importance whatever and that they should be left out of the evaluation. In this way, a specific evaluation can be extracted from a general quality model (ibid.).

In the first period of EAGLES work, it was thought important to validate the theoretical ideas built on ISO work by designing practical evaluations of relatively simple language engineering products. Thus, a rather thorough evaluation of spelling checkers were designed and carried out, and preliminary work was done on designing evaluations for grammar checkers and for translation memory systems. It was both salutary and sobering to be forced to realize the sheer amount of meticulous attention to technical detail
required in order to construct rigorous evaluations even of systems whose
technology is relatively simple (Hovy & King, 2001: 5).

Hovy (1999) maintains that the EAGLES Report (1996) lists the validity of measures to users, the types of attributes, and the kinds of tests that can be applied. The study is extremely thorough and thought-provoking. It provides examples of evaluations of aspects of various types of MT systems. In all, “it is something anyone interested in the subject of MT evaluation has to read”. However, it does not yet satisfy the criterion of simplicity. There is no direct link between the user’s task and any evaluation measure. As a result, the user does not know which of the measures to apply, and how much importance to give to each one. Further, the sheer number of evaluation measures is rather overwhelming. The user is almost required to become an expert in MT evaluation before being able to make a decision.

2.5.1.3 The EAGLES 7-Step Recipe
The second phase of the EAGLES initiative covered 1995 and 1996. Work on evaluation during this phase was essentially limited to consolidation and dissemination of previous work. During this time, the EAGLES methodology was used outside the project to design evaluations of a dialogue system and of a speech recognition system, as well as a comparative evaluation of a number of dictation systems (Blasband, 1999). The designers of these evaluations provided useful feedback and encouragement to the group. Also, during the second phase, the EAGLES group comes into close contact with the ISO/IEC work on the evaluation of software in general (For more details, see EAGLES, 1999: 6-17; Hovy & King, 1999: 6-10).
Recently, the importance of the usability of a product has been recognized and stressed. The *EAGLES Evaluation Working Group* (1999) proposed a general framework for evaluation following ISO quality model. Their report emphasized the importance of quality in use as well as quality of a product. They defined quality in use as the user’s view of the quality of a system containing software, claiming that it is measured in terms of the results of the use of software like an MT system that is in terms of effectiveness, productivity, and satisfaction of users (Koh et al., 2001).

In the *EAGLES Report* of 1999, the 7 major steps necessary to carry out a successful evaluation of language technology systems or components are discussed in detail with exemplification (*EAGLES Report*, 1999). Here, only a brief overview of these steps is presented:

The 7-Step Recipe includes:

1. Why is the evaluation being done?
   - What is the purpose of the evaluation? Do all parties involved have the same understanding of the purpose?
   - What exactly is being evaluated? Is it a system or a system component? A system in isolation or a system in a specific context of use? Where are the boundaries of the system?

2. Elaborate a task model:
   - Identify all relevant roles and agents.
   - What is the system going to be used for?
   - Who will use it? What will they do with it? What are these people like?

3. Define top level quality characteristics:
   - What features of the system need to be evaluated? Are they all equally important?
4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3:
   - For each feature which has been identified as important, can a valid and reliable way be found of measuring how the object being evaluated performs with respect to that feature? If not, then the features have to be broken down in a valid way, into sub-attributes which are measurable. This point has to be reported until a point is reached where the attributes are measurable.

5. Devise the metrics to be applied to the system for the requirements produced under 4:
   - Both measure and method for obtaining that measure have to be defined for each attribute.
   - For each measurable attribute, what will count as a good score or an unsatisfactory score given the task model (2)? What are the cut off points?
   - Usually, an attribute has more than one sub-attribute. How the values of the different sub-attributes are combined to a value for the mother node in order to reflect their relative importance (again given the task model)?

6. Design the execution of the evaluation:
   - Develop test materials to support testing of the object.
   - Who will actually carry out the different measurements? When? In what circumstances? What form will the end result take?

7. Execute the evaluation:
   - Make measurement.
   - Compare with the previously determined satisfaction ratings.
   - Summarize the results in an evaluation report, cf. Point 1.
In short, evaluation is needed to know if a given system fulfils its promises; however, it also serves other side purposes. It brings out software’s characteristics better than other methods such as, reports, software manuals and training courses. It focuses on software pros and cons. It also allows users to compare different software products on similar bases. So, in order to carry out a successful evaluation, the above-mentioned evaluation steps need to be systematically followed, although in real life, the situation would be more complex and requires much more details.

### 2.5.1.4 The ISO/IEC Standards for Software Evaluation

#### 2.5.1.4.1 A Growing Set of Standards

According to what is mentioned in the *EAGLES Report* of (1996) and the *EAGLES Report* of (1999), ISO together with IEC have initiated in the past decade an important effort towards the standardization of software evaluation. In 1991, appeared the ISO/IEC 9126 standard (*ISO/IEC-9126, 1991*), a milestone that proposed a definition of the concept of quality, and decomposed software quality into six generic quality characteristics.

Subsequent efforts led to a set of standards, some still in draft versions today. It appeared that a new series was necessary for the evaluation process, of which the first in the series (*ISO/IEC-1459 - 1298, 2001, Part 1*) provides an overview. The new version of the ISO/IEC 9126 standard will finally comprise four inter-related standards: standards for software quality models (*ISO/IEC-9126-1, 2001*), for external, internal and quality in use metrics
Regarding the 14598 series and volumes subsequent to ISO/IEC 14598-1, the focus is on the planning and management (14598-2) and documentation (14598-6) of the evaluation process, and the application of the generic organization framework to developers (14598-3), acquirers (14598-4) and evaluators (14598-5) (For more details, see the EAGLES Report, 1999: 5-17).

2.5.1.4.2 The Definition of a Quality Model
This subsection situates the proposal presented by Hovy, King and Popescu-Belis (2002) for MT evaluation within the ISO/IEC framework. According to ISO/IEC 14598-1 (1998/2001, Part 1: 12), the software life cycle starts with an analysis of user needs that will be answered by the software, which determine in their turn a set of specifications. From the point of view of quality, these are the external quality requirements. Then, the software is built during the design and development phase, when quality becomes an internal matter related to the characteristics of the system itself. Once a product is obtained, it is possible to assess its internal quality, then the external quality, i.e., the extent to which it satisfies the specified requirements. Finally, turning back to the user needs that were at the origin of the software, quality in use is the extent to which the software really helps users fulfil their tasks (ISO/IEC 9126-1, 2001: 11).

Hovy, King and Popescu-Belis (2002) point out that quality in use does not follow automatically from external quality since it is not possible to predict all the results of using the software before it is completely operational. In addition, for MT software, there seems to be no straightforward link, in the conception phase, from the external quality requirements to the internal structure of a system. Therefore, the relation between external and internal qualities is quite loose.
Following mainly (ISO/IEC 9126-1, 2001), software quality results from six quality characteristics:

1. Functionality.
2. Reliability.
3. Usability.
4. Efficiency.
5. Maintainability.
6. Portability.

These characteristics have been refined into software sub-characteristics that are still domain-independent (ISO/IEC 9126-1, ibid.). These form a loose hierarchy, but the terminal entries are always measurable features of the software, i.e., attributes. Following (ISO/IEC-14598, 1998-2001, Part 1), a measurement “is the use of a metric to assign a value (i.e., a measure, be it a number or a category) from a scale to an attribute of an entity” (Hovy and King, 2001: 8).

The six top level quality characteristics are the same for external as well as internal quality. The hierarchy of sub-characteristics may be different, whereas the attributes are certainly different, since external quality is measured through external attributes (related to the behaviour of the system) while internal quality is measured through internal attributes (related to intrinsic features of the system) (ibid.).

Finally, quality in use results from four characteristics: effectiveness, productivity, safety and satisfaction. These can only be measured in the operating environment of the software, thus seeming less prone to standardization (See Daly-Jones et al., 1999 and ISO/IEC 9126-1, 2001).

2.5.1.4.3 Stages in the Evaluation Process
The ISO/IEC standards outline the evaluation process, elaborating on a proposal already present in the first ISO/IEC 9126, towards a more general, hence abstract model. The five consecutive phases of the evaluation process according to *(ISO/IEC-9126, 1991: 6)* and *(ISO/IEC-14598, 1998-2001, Part 5: 7)* are:

1. Establish the quality requirements (the list of the required quality characteristics).
2. Specify the evaluation (specify measurements and map them to requirements).
3. Design the evaluation, producing the evaluation plan that documents the procedures used to perform measurements).
4. Execute the evaluation, producing a draft evaluation report.
5. Conclude the evaluation.

As Hovy and King (2001:10) indicate, during specification of the measurements, each required quality characteristics must be decomposed into the related sub-characteristics, and metrics must be specified for each of the attributes arrived at in this process. More precisely, three elements must be distinguished in the specification and design processes; these correspond to the following stages in execution:

- Application of a metric.
- Ratings of the measured value.
- Integration (assessment) of the various ratings.

It must be noted that (a) and (b) may be merged in the concept of ‘measure’, as in ISO/IEC 14598-1, and that integration (c) is optional. Still, at the level of concrete evaluations of systems, the distinction, advocated also by *EAGLES Report* (1996) seems particularly useful: to evaluate a system, a metric is applied for each of the selected attributes, yielding as a
score a raw of intrinsic score; these scores are then transformed into marked or rating levels on a given scale; finally during assessment, rating levels are combined if a single result must be provided for a system (Hovy, King and Popescu-Belis, 2002).

A single final rating is often less informative, but more adapted to comparative evaluation. However, an expandable rating, in which a single value can be decomposed on demand into several components, is made possible when relative strengths of the component metrics are understood. Conversely, the EAGLES methodology considers the set of ratings to be the final result of the evaluation (EAGLES, 1996: 15).

To conclude this section on the ISO/IEC standards, it is quite apparent that they have been conceived as an abstract framework that suits the needs of many communities that develop or use software. Hovy, King and Popescu-Belis (2002) particularize this framework to MT evaluation, starting with an essential factor that influences the choices made among quality characteristics. This makes visible the link with previous standardization efforts in MT evaluation.

2.5.1.5 The ISLE MTE Framework
ISLE is the latest in a series of projects under the successful EAGLES initiative. It extends hitherto European-based EAGLES work within the European-American International research cooperation framework, set up as a result of two years of joint preparatory work towards an International Human Language Technology (HLT) standards oriented initiative.

The overall goals of ISLE are to support HLT and national projects and HLT industry in general by developing, disseminating and promoting widely agreed and urgently needed HLT standards and guidelines. In an effort to
develop a more systematic MT evaluation methodology, recent work in the EAGLES and ISLE Projects has created a framework of characteristics in terms of which MT evaluations and systems, past and future, can be described and classified. The results of this work are intended to help three types of user: 1.) people who need guidance in choosing an MT system, 2.) those who want to compare various MT systems and 3.) those who want to design a new MT system or upgrade an existing one (ISLE Annual Report, 2001) and (LREC Workshop, 2002).

The focus of work on evaluation is on methods and metrics for MT. This has involved investigation of the various published evaluations of MT systems that have been carried out since 1979. A second version of a specific framework for classifying MT evaluation has been elaborated, illustrating how the current state of the ISLE evaluation methodology can be applied (See 2.5.1.5.2 below).

2.5.1.5.1 Origins and First Version

While the EAGLES Projects witnessed the development of a formalization of NLP software evaluation, it was only in the ISLE follow-up project that a systematic application to MTE was stated. The fundamental idea of the application was to list, in a hierarchical order, all the features pertaining to the quality of MT software and the main metrics associated to them.

As Popescu-Belis, Manzi and King (2001) state, the first draft of this taxonomy was developed at the Information Science Institute by Eduard Hovy and Elena Filatova in 2000, based mainly on previous work by Eduard Hovy presented at an EAGLES Workshop (Hovy, 1999). This proposal contained two parallel levels of classification, one containing measures related to the purpose of an MT system, and the other containing measures
related to the translation process itself. As the author acknowledged, these were only example taxonomies, and did not represent a final or complete classification. A significant choice was further made to design a hypertext document and make it available over the Internet. This first version is available at: www.isi.edu/natural-language/mteval.

The introduction to this draft particularized the classification principle, stating that each characteristic is sub-divided into more detailed features by advocating a strong initial dichotomy. At the top level of the taxonomy, there is only a single item to consider, namely the grand unified evaluation score; one level lower, there are only two items to consider, namely the score for how well the system achieves the user’s purposes and how well the system performs its internal operations. The first-draft provides three different parallel [sub] taxonomies: a.) user purpose, b.) application process and c.) general software characteristics. In contradistinction, the second draft grants a special role to the first level articulation between user needs and system characteristics (Popescu-Belis, Manzi and King, 2001).

2.5.1.5.2 Key Principle of the Second Draft

The first draft was discussed during a workshop at the AMTA Conference Evaluation Workshop (Reeder and Hovy, 2000). In an unpublished document, Maghi King synthesized contributions of the participants to the discussions, and proposed a significant restructuring of the taxonomy while preserving most of its contents, i.e., the lower level categories and the individual characteristics (Popescu-Belis, Manzi and King, 2001). The central point of this proposal was the articulation for evaluation purposes namely the user’s needs and the system’s characteristics. More precisely, taking into account the central role that the user of a system plays
with respect to evaluation, the classification was divided into two complementary sections (ibid.):

The first part relates user needs to characteristics of the system. In other words, it is a repertoire of possible tasks for an MT system, user profiles, document types and qualities, organized hierarchically. Each item is refined using several alternative or complementary lower-level items. The entry for each item, taken from the first version, describes the system characteristics that are relevant for the item, and should hence be evaluated.

The second part is complementary to the first, since it contains, for each system characteristics, one or more metrics that have been proposed to quantify its quality level. These metrics were extracted, already in the first draft, from the MT Literature. Along with the metrics, each entry provides a definition, comments and references for the corresponding characteristic.

Thus, the evaluators of an MT system should first use the first part to identify, for the desired task and user profile, the relevant characteristics, then attempt to evaluate each of these using the metrics described in the second part. It is worth noticing, here, that a hypertext taxonomy has been developed, and enhanced in the second draft (2001) available at: http://www.issco.unige.ch/projects/isle/taxonomy2.

2.5.1.5.3 A Framework for the Evaluation of Machine Translation in ISLE (2003)

In order to be of use, the classifications of evaluation methods and measurements must be comprehensive, up-to-date and accessible. On the basis of this, the ISLE group has updated their taxonomy in February 2003,

The FEMTI backbone is made of two classifications or taxonomies. The first one is intended to be used in determining what the user’s needs are. The second concerns quality characteristics of MT systems which are potentially of interest. Certain modifications have been made to the second version of the ISLE taxonomy in its recent version of taxonomy 3. The schema below gives a general view of the contents of the two taxonomies. The first one enumerates non exclusive characteristics of the context of use grouped in three complementary parts (task, user, input). The second one develops the quality model, and its starting point in the six ISO/IEC quality characteristics. As the link between internal features and external performance is not yet complementary clear for MT systems, the internal attributes are structured in a branch separate from the six ISO/IES characteristics which are measured by external metrics (FEMTI, 2003) and (Popescu-Belis, 2003 b). For lack of space, the hierarchies below represent a brief snapshot of the actual state of this proposal:

**COMPACT CLASSIFICATION**

1 Evaluation requirements

- 1.1 The purpose of evaluation
- 1.2 The object of evaluation
- 1.3 Characteristics of the translation task
  - 1.3.1 Assimilation
  - 1.3.2 Dissemination
  - 1.3.3 Communication
• 1.4 User characteristics
  o 1.3.1 Machine translation user
  o 1.3.2 Translation consumer
  o 1.3.3 Organizational user

• 1.5 Input characteristics (author and text)

2 System characteristics to be evaluated

• 2.1 System internal characteristics
  o 2.1.1 MT system-specific characteristics
  o 2.1.2 Translation process models
  o 2.1.3 Linguistic resources and utilities
  o 2.1.4 Characteristics of process flow

• 2.2 System external characteristics
  o 2.2.1 Functionality
    ▪ 2.2.1.1 Suitability
    ▪ 2.2.1.2 Accuracy
    ▪ 2.2.1.3 Wellformedness
    ▪ 2.2.1.4 Interoperability
    ▪ 2.2.1.5 Compliance
    ▪ 2.2.1.6 Security
  o 2.2.2 Reliability
  o 2.2.3 Usability
  o 2.2.4 Efficiency
  o 2.2.5 Maintainability
  o 2.2.6 Portability
  o 2.2.7 Cost

It is this model of Taxonomy 3 (FEMTI, 2003) which is going to be adopted in the present evaluation of the MT systems in this study. Within this framework, the criteria of the MT systems to be selected and measured and the appropriate measures to be used in the present work will be identified and explained in detail in the next chapter.
2.5.1.5.4 Analyzing the Behaviour of Measures

Since the above-mentioned proposed taxonomy (See 2.5.1.5.3) gathers numerous quality attributes and metrics, there are basic aspects of MT that may be rated through several attributes, and each attribute may be scored using several metrics. Actually, this uncomfortable state of affairs calls for investigation.

According to Hovy and King (2001: 29-31); and Hovy, King and Popescu-Belis (2002) the measure:
1. Must be easy to define: clear and intuitive
2. Must correlate well with human judgements under all conditions, genres, domains, etc.
3. Must be ‘tight’, exhibiting as little variance as possible across evaluators, or for equivalent inputs.
4. Must be cheap to prepare (i.e., not require a great deal of human effort for training data or ideal examples).
5. Must be cheap to apply.
6. Should be automated if possible.

In this respect, few evaluators have both bothered to try someone else’s measures and correlate the results. However, there are some methodological studies of this kind. Papineni et al., (2001), for instance, compared the score given by BLEU (an algorithm), with human judgements of the ‘fluency’ and ‘fidelity’ of translations. They found a very high level of agreement, with correlation coefficients of 0.99 (with monolingual judges) and 0.96 (with bilingual ones).

Another important matter is ‘inter-evaluation agreement’, reported on by most careful evaluations. There is still lack of guidelines for formulating the
instructions for evaluators, and no idea how variations would affect systems’ scores. Similarly, it is unknown whether a 3-point scale is more effective than a 5- or 7-point. So, experiments are needed to determine the optional point between inter-evaluator consistency (higher on shorter scale) and evaluation informativeness (higher on longer scale). Still, another important issue is the number of measure points required by each metric before the evaluation can be trusted, a figure that can be inferred from the confidence level of past evaluation studies (Hovy and King, 2001:29-30) and (Hovy, King and Popescu-Belis, 2002).

The ISLE research working group is now embarking on the design of a programme that will help address these questions. Their very ambitious goal is to know, for each criterion in the taxonomy, which measure(s) are most appropriate, which metric(s) to use for them, how much work and cost is involved in applying each measure, and what final level of score should be considered acceptable (or not). Armed with this knowledge, a would-be evaluator would be able to make a much more informed selection of what to evaluate and how to go about it. It is this taxonomy that is going to be followed in the present study since it is the most recent and comprehensive (confirmed in Hovy, 2002a).

Finally, the ISLE Research Working Group aims at investigating whether their work (described here) can be extended to other fields. Some previous experience has shown that it applies relatively (straightforwardly) to some domains, such as dialogue systems in a specific context of use. However, as the systems to be evaluated grow more complex, the contexts of use become potentially almost infinite. Their belief is that the basic ISO notion of building a quality model and associating appropriate metrics to it should carry over to almost any application.
2.6 New Trends and Future Work in MTE

Although substantial progress has been made towards the overall goal of defining a general framework for the design of evaluation methodologies (e.g. EAGLES, 1996 & 1999); developing practical and objective MTE criteria (e.g. JEIDA, 1992); developing international standards related to evaluation (e.g. ISO/IEC, 1991); and building a taxonomy of features for classifying measures of evaluation, much remains to be done. In general, work regarding MTE in the previous period has led to the identification of areas when very little indeed has been done in the large MT community which should be given priority in follow-up work.

For EAGLES group, obvious directions for further work include applying their framework to other kinds of evaluation (as they have concentrated only on ‘adequacy’ evaluation), and to other areas, perhaps even outside the strict domain of language engineering. In addition, very little has yet been done on identifying and formalizing the needs of classes of users. Thus, they intend to go beyond the bounds of language engineering products to touch the concerns of industry as a whole and potentially provide insights useful to research policy and research and development projects (EAGLES, 1996: 57).

Hovy, King and Popescu-Belis (2002) indicate that while it is possible to continue refining the taxonomy (they have built), collecting additional references and classifying additional measures, the most pressing work regarding the refinement of the taxonomy “is only now being started”. This work is but the first step toward a more comprehensive and systematic understanding of MTE in all its complexity.
New trends in MTE have focused on ways to automate MTE, without relying on time-intensive, human-subject elicitations. These studies, for example, Hirschman, et al., (2000); Jones and Rusk (2000); Niessen et al., (2000: 39-45); White (2000); Papineni et al., (2001); Reeder et al., (2001); Papineni (2002); Popescu-Belis, King and Benantar (2002) and others have raised the possibility that automatic measurements of certain attributes or a set of attributes of MT output might be extrapolated to predict measures germane to MT itself. Several automatic measures for MTE have been proposed and computational tools to carry them on effectively are now available. One major advantage of this is that evaluation processes are greatly accelerated.

Hovy, King and Popescu-Belis (2002) confirm that MTE is simply a special, although rather complex case of software evaluation in general. The dream of a magic test that makes everything easy-preferably an automated process always remains. For although the existence of a quick and cheap evaluation measure is enough for many people, it still does not cover more than a small portion of the aspects of MT that people have wished to measure.

A recent trend in MTE looks at the evaluation of language learners as a source for techniques in MTE. Language learner evaluation has had similarly checkered career-methods for accurately measuring language competence that have changed to reflect trends of pedagogy and computing ability. Language learner evaluation research, however, has developed some simple tests which have shown strong correlations to language ability and are good indicators of language competence. These are exactly the kinds of measures which are being sought for in MTE (Vanni and Reeder, 2001).
The direction of these endeavours seems to be toward streamlining the evaluation process and equipping users with tools for carrying out their own evaluations of MT systems which are tailored to what the user requires from the MT system output. One feature of any such test will be a description of what linguistic features the system can handle reliably. Another possibility is to look at second language acquisition errors. Research in second language acquisition and also cognitive skills development provides us with a potential model for identifying a constellation of such features useable diagnostically to characterize the performance of a system (Conner – Linton, 1995: 99-115).

2.7 Formal and Functional Approaches to MTE

Early evaluations of MT focused largely on analysis of the output and on cost–effectiveness of the throughput. More recently there has been awareness that a system’s value depends on the use to which it will be put. The focus has shifted to its suitability for the application in question. Consideration must be given to a broad range of factors in the system’s environment before a judgement can be made about its overall effectiveness. These factors are emphasized differently by developers, managers, translators, end-users of the product.

Formal evaluations of MT deal with parameters that are susceptible to measurement. Most commonly they focus on the machine output. Although, there is no way that these studies can give a total picture of how MT works in a given environment, still they can be meaningful when undertaken in perspective. For example, studies of output quality have informative value
for these end-users who will be dealing with the raw product directly, and they should not be discounted for such purposes when combined with other criteria. Formal results alone, however, do not provide an adequate basis on which conclusions can be drawn about the effectiveness of MT. Far more useful is a functional approach in which account is taken of all the tasks and other factors involved in meeting the purpose that the system is intended to serve in the particular environment. In actual practice, it is to be expected that most MT evaluations will include elements from both the formal and the functional approaches (Vasconcellos, 1988: 203-205).

The most famous evaluation of MT was the one that led to the ALPAC Report in (1966) by the U.S National Academy of Sciences. As almost everyone involved in MT knows, the results of this study were to cast a shadow on MT development in the U.S. that persists until the present day. The approach was basically formal. The ALPAC Committee focused on the raw machine output and undertook to evaluate the product in terms of two major characteristics of translation, i.e., ‘intelligibility’ and ‘fidelity’ to the sense of the original text. As a reaction to ALPAC, other approaches began to be sought for the evaluation of MT. They made an assessment to both the quality of the output and the cost of throughput (e.g. evaluation of SYSTRAN by Van Slype, 1979a). Such error analysis of MT output was then, and continues to be, a fairly frequent exercise, since it is the first possibility that comes to mind when one is initially confronted with translations produced by machine (ibid : 206-209).

In general, there are other parameters that can be treated formally in addition to quality of output. For instance, some studies measure the time required to bring the machine–translated text up to the desired level of
quality. Others consider the speed of revision as a parameter, combining it with a large set of criteria of the output.

As Vasconcellos (1988: 210) maintains a fully functional evaluation of an MT system will take into account not only end-user satisfaction, but also the concerns of management and of the post-editing translators. Moreover, since a dynamic environment includes interaction with the developer, consideration should also be given to the needs, priorities and constraints of the latter. But, more than anything else, what must be determined is whether or not the MT system has the capacity to grow and that the human principals who will be interacting with it will be able to make effective contributions to that growth. An MT system’s potential hinges on four main factors: a.) the dictionary structure, b.) the daily working environment, c.) the translators who use it, and d.) the ongoing support provided by the developer. The functional evaluation should be designed to determine the existence of these conditions. If these can be assured, then the system can be counted on to flourish in its new environment.

All in all, there is no single ideal model for MTE. But, even though there is no single right way to go about an evaluation, it is safe to say that formal data should be eyed from their inherently limited perspective and that priority should always be given to the functional factors that shape the future of the system.

2.8 Evaluation for MT Stakeholders

How can we evaluate at all? The answer lies in controlling the factors that can be controlled and optimizing the control of those that cannot be captured completely. The largest best control is to make the common-sense distribution among the different things that different people need to know
about MT systems; in other words, that no one evaluation method will fit all needs. It is obvious that the most important people in the world of MT are translators. However, there are several other groups of people who have a “stake” in the success of one more aspect of MT. White (in Somers, in press) divides these into:

1. **End-users**
   
i. **Translators:** Who need MT systems which are easy to access and use, compatible with their computerized environment and work processes. The system must enable translators to make the best possible use of their expertise and experience, to increase the quality of translations they can do.

   
   ii. **Translation editors:** In the professional translation environments where post-translation editing is part of the work process, the editors have the specialized requirement of making sure that translations are both accurate and consistent with other translators’ work in the same document set. Current MT systems impose additional editorial requirements on these people. These requirements should be easy to meet, and the system as well should make the pre-existing job of quality control, version control, and consistency easier as well.

   
   iii. **Monolingual information consumers:** These are people, who need information at one time or another that is comprehensive, relevant and timely, with little or no regard for the language or its origin. Here, the work of the MT system in the overall flow of automatic information processing should be transparent to the information consumer.

   
   iv. **Office automation users:** People are more and more accustomed to inter-operating suites of applications. The days of single purpose, “turnkey” computes are long gone. If MT systems are operated in the
course of work, these people should expect them to accept input from other office automation applications, and return output compatible with other applications in the office automation suite.

2.) Managers

i. Operational managers: These need to know if an MT system will work in the environment of their translator employees, given the environment, requirements, etc. The operational manager needs to know whether MT will actually improve the performance of the translation department.

ii. Procurement managers: The people responsible for purchasing systems need to know whether the system requires equipment or connectivity that the department presently does not have or requires special licenses or usage cost that might have to be taken into account. They also would like to know whether the company that provides a system is sufficiently sound and viable to provide support and upgrades.

3.) Developers

i. Researchers: There are many types of research with different objectives and at different levels of maturity. However, a common need in research is to know whether a particular approach actually matches the hypothesis for its success. Other issues in research have to do with the extensibility of a translation approach beyond a particular set of phenomena into the infinite world of real language use.

ii. Productizers: The people who take the fruits of research and attempt to make a marketable product need to know whether the conceptual prototype can ramp up to meet the needs of real use, and can fit into a real automation environment.
4.) **Vendors:** They need to know whether an MT system they wish to sell is robust and extensible enough to fit into a variety of different settings, i.e., if the demand is sufficient to justify the marketing and support investment.

5.) **Investors:**
   - **Research organizations:** Organizations such as government agencies that sponsor research need to know whether sufficient progress is being made to demonstrate the research hypothesis, and that results that do appear are not artifacts of extraneous effects.
   - **Venture capitalists:** People interested in investing in high technology in general need to know whether MT is a worthwhile endeavor. Here, they need to know whether the technology is in fact viable, whether the companies trying to develop and market it are stable, and what the future trends for demands and state of the art will be.

### 2.9 Types and Purposes of MTE

As it is mentioned above, the different responsibilities and obligations of each of the stakeholder groups means that each group needs to know different things about MT. The end-user in a translation environment does not need to know what the cost of a system is in order to do his job. Nor does the end-user have to know where in the analytical engine pronominal reference is handled, unless the user interface is rather primitive. The investigator for a particular scientific approach to MT is unlikely to be concerned, at first, about whether a system that will someday incorporate the results of their findings will run efficiently on a conventional desktop computer. Thus, each stakeholder’s need for information must be covered by a particular, pertinent, set of evaluation types.
The researcher of the present study will layout a descriptive model of evaluation types and discuss each type along with some of the fundamental issues that arise for each type of MTE. The classification of evaluation types is based on Hirschman and Thompson (2000); FEMTI (2003) and White (in Somers, in press).

2.9.1 Feasibility Evaluation

According to FEMTI (2003:1-2) and White (in Somers, in press), a feasibility study is an evaluation of the possibility for a particular approach whether it has any actual potential for success after further research and implementation. Feasibility evaluations provide measures of interest to researches and the sponsors of research. The characteristics that a feasibility evaluation typically tests are functionality attributes such as the coverage of sub-problems particular to a specific language pair and the possibility of extending to more general phenomena (changeability).

For instance, if there is a claim that a particular linguistic or computational and/or implementation approach will do translation somehow “better” than existing approaches. So, in addition to showing fundamental coverage of the contrastive issues, two other things about this approach must be shown: a.) that its good points facilitate the coverage results and b.) its bad points do relatively little harm.

2.9.2 Internal Evaluation

Internal evaluation occurs on a continual or periodic basis in the course of research and/or development. Here, the question is whether the components of an experimental, prototype or pre-release system work as they are intended. The particular items covered in such an evaluation will vary with
the maturity of the system being evaluated of course, and thus provide results of interest to researchers, research sponsors, developers, and vendors (FEMTI, 2003: 2-3 and White in Somers, in press).

As with the feasibility test, it is to be able to show that the fundamental constructive phenomena of the language pair can be covered. But, we need to have a standard set of test materials for interactive testing\(^2\) to show some other attributes as well, namely that the system we are developing, or bringing to market, or adapting to our own user environment, is improving. For instance, we need to show that as we add grammar rules or dictionary entries, the system translates the things we are trying to improve better than it did, and does not suddenly fail to do something it used to do. Moreover, in this type of evaluation, we need to show that the implementation of our approach can also extend beyond certain very focused patterns into the language text that will actually occur in production. So internal evaluation typically handles both: a.) the patterns for regression testing against specific phenomena, and b.) actual text for determining extensibility (EAGLES, 1996:18 and White in Somers, in press).

Fundamentally, we have some input, some output, and a design in which the system’s components do distinct things that come together in the intended way to produce the intended output. How we regard these elements in evaluation gives us different, and equally useful, views of internal performance and predictable extensibility (White in Somers, in press).

### 2.9.3 Black-box Vs. Glass-box Evaluation

A distinction is often drawn between so-called glass-box and black-box evaluation. This concerns the relationship of the input and output. The black-box view is a look at the input and output without taking into account
the mechanics of the translation engine. The glass-box view looks inside the translation engine to see if each of its components did what was expected of them in the course of the translation process (*EAGLES*, 1996: 68ff; Hirschman & Thompson, 2000; *ISLE*, 2003, 3-4 and White in Somers, in press).

White in Somers (ibid.) indicates that there are advantages to each. The black-box view is portable, i.e., the method and measures are external to the design and philosophy of anyone system. It is more amenable to comparisons of systems and to determining the current language of a particular system. The glass-box view helps to determine the extensibility of coverage of the system, by being able to tell whether and how well the designed processes perform their functions. Did, for instance, the transfer rules correctly move a prepositional phase to the right position, or did an apparently correct result come from a fortuitous default?

2.9.4 Diagnostic Evaluation

This type of evaluation has the purpose of discovering why a system did not give the results it was expected to give. Typically performed by a researcher developing a prototype system, such an evaluation is almost exclusively concerned with functionality characteristics and will also often make use of internal metrics based on the intermediate results the system produces. Unlike the other types of evaluation distinguished in this classification, it is a glass-box evaluation (Hirschman & Thompson, 2000 and *FEMTI*, 2003: 3-4).

2.9.5 Declarative Evaluation
This evaluation is the heart of the matter for the casual observer. It addresses the question whether a system translates well, i.e., the degree to which it has the attributes of fidelity and intelligibility. This evaluation type is clearly of particular value to investors, end-users, vendors and managers, but also to developers (FEMTI, 2003: 4).

In declarative evaluation, the methods used are expected to be very much like those of internal evaluation—coverage of linguistic phenomena and handling of samples of real text, to name two obvious methods. Here, more interest is in what the system can currently manage than in what its extensibility potential is; therefore, it may be more likely to look at black-box views (White in Somers, in press).

Declaration evaluations generally test for the functionality attributes of intelligibility (how fluent or understandable it appears to be) and fidelity (the accurateness and completeness of the information conveyed). Three principal methods are used in this type of evaluation: a.) analysis of errors, b.) rating of ability to do a task as a result of the output and c.) rating of intuitive judgement of the “goodness” of the translation (FEMTI, 2003:4 and White in Somers, in press).

2.9.6. Adequacy Evaluation

As speech and natural language processing systems move out of the laboratory and into the market, it is becoming increasingly important to address the legitimate needs of potential users in determining whether any of the products on offer in a given application domain are adequate for their particular task, and if so, whether any of them are obviously more suited than the others (Hirschman and Thompson, 2000).

The term formative evaluation is used in the field of human-computer interaction to refer to a collection of evaluation methodologies more closely
related to both adequacy evaluation and to diagnostic evaluation in our terms. The goal of formative evaluation is to provide diagnostic information about where a given system succeeds or needs improvement, relative to its intended users and use. The role of formative evaluation is to influence and guide system design, as opposed to performance evaluation or summative evaluation, which rates systems relative to each other, or relative to some gold standard such as human performance (ibid.).

2.9.7 Performance Evaluation
This is the measurement of system performance in one or more specific areas. In considering any attempt at performance evaluation, three levels of specificity can be distinguished: a.) criteria, i.e., what is interesting in evaluation, in the abstract: precision, speed, error rate, b.) Measure, i.e., which specific property of system performance is reported in an attempt to get at the chosen criterion: ratio of hits to hits plus misses, second to process, percent incorrect, c.) Method, i.e., how the appropriate value for a given measure and a given system is determined: typically some form of concurrent or post-analytic measurement of system behaviour over some benchmark task. It should be clear that the distinction between criterion, measure, and method is not hard and that in any given case, the three are interdependent (ibid.).

2.9.8 Operational Evaluation
As White in Somers (in press) points out, operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use. The primary factors include the
cost-benefit of bringing the system into the overall process (cost). Operational evaluations answer the question “Is it worth it?” Issues like common platforms and operating systems are germane here. End users and their managers need these evaluations, and thus investors and vendors must be attractive to the operational factors.

A variety of issues are considered here, including such things as software and hardware compatibility with the incumbent office automation system (interoperability). However, the more fundamental question to ask for operational use is whether the MT system enhances the effectiveness of the downstream task, or whether the end-to-end process is better off without it (FEMTI, 2005: 5).

2.9.9 Usability Evaluation

According to White (in Somers, in press) the purpose of usability evaluation is to measure the ability of a system to be useful to people who are actually going to use it. ISO 9126 talks of “quality in use” characteristics, which are the combination of other characteristics which will allow a user to perform specific task effectively, safely and with satisfaction.

Usability evaluation is a domain in its own right, which involves kinds of testing such as scenario and laboratory testing common to many kinds of software product. It is often undertaken by the manufacturers of products before the product is launched on the market. It falls outside the scope of the current classification. However, much information about usability can be found by consulting the European Usability Support Centers home page (FEMTI, 2003:6).

2.9.10 Comparison Evaluation
Comparisons measure some attributes of a system against the same attributes of other systems. Thus, the methods of comparison are the same as the methods of the other evaluation types, applied among several systems. This is of obvious benefit to purchasers of systems and investors in system development and production. The purpose of comparison evaluations is to determine the best system, best implementation, or even the best theoretical approach for meeting current or future needs. It appears that comparison evaluation can measure the same attributes as the other types (e.g. feasibility, internal and operational) of evaluation studies. Depending on what we are comparing, it has all of the properties of any of these other types, except that in each case we are holding the measurements of one against the same measurements of another. For instance, counting errors can be used to compare systems by errors produced, with all other factors optimally controlled (White in Somers, in press).

In fact, comparison evaluation can use any of the same methods as the other types. The caution is, of course, that the methods have to make sense for what it is being compared. So, the methods of the JEIDA study can be readily used to determine which system of several possible candidates to select, but cannot tell which of two linguistic approaches gives the best results for prepositional-phrase attachment, for instance (ibid.).

The above discussion has made it obvious that most of the measures that might be attempted on MT are subjective in nature, whether for good or ill. This implies at the very least that it takes some time as people do not do analytical things very quickly compared to computers. Therefore, some ways to automate measures germane to several types of MT evaluation have been applied recently.
2.10 The Object of Evaluation

It is important to determine what context of use is to be taken into consideration in the evaluation. Whilst it is impossible to give a detailed breakdown of all possible contexts of use, an example may help to clarify what is meant here (FEMTI, 2003: 6):

e.g., Imagine an MT system as one component in a system whose overall purpose is to retrieve and present to the user information on railway timetables, accepting voice input and producing voice output. The boundaries could be taken as:

- a. actual speakers receiving actual information;
- b. the MT system receiving artificially constructed transcribed input on one side and the spoken output on the other;
- c. the MT system receiving artificially transcribed input and producing as output the input to the speech synthesizer;
- d. the MT system receiving artificially transcribed input and producing a query to the timetable query system;
- e. the MT system receiving information from the timetable query system generating from it the input to the speech synthesizer.

It is clear that many other possibilities exist, especially if the boundary is further widened to include different types of end users and different places where the system may be installed.

The object of evaluation may be: a.) an MT system considered as a whole, b.) a component of an MT system and c.) an MT system considered as a component of a large system (ibid: 7).

In the present study, the object of evaluation involves testing certain components of an MT system in terms of selected system external
characteristics. This is germane to the glass-box/black-box distinction. Here, concentration will be on the black-box view regarding the relationship of the input and output without taking into account the mechanics of the translation engine (Guessoum, 2002b). According to FEMTI (2003: 6), when an MT system is considered as a whole, this refers to the typical case for the instantiation of the entire taxonomy in its current state. Moreover, when an MT system is considered as a component of a large system, the evaluation critically depends on the large application, and the metrics will be those of the “upstream” and “downstream” processes. This aspect of MTE is not dealt with in any detail in the current version of this taxonomy (i.e., FEMTI, 2003). So, these two considerations of the object of MTE are out of the scope of the present work.

2.11 MT Systems

2.11.1 Background

In general, MT systems have succeeded in increasing translation productivity for many years, and progress in this area continues. MT remains one of the very few areas in the practice of NLP where the relative successes or failures of systems have been measured in objective, even commercial, terms. Regarding NLP theory, high-quality translation is certainly one of the most challenging possible tests, and MT therefore offers a rich matrix for continued exploration and experimentation.

Since the middle of the 1990s, there has been a rapid increase in the number and variety of translation systems available, in the form of stand-alone software for ‘automatic’ translation, computer-aided translation systems for large corporations, translator workbenches, translation memory systems, on-line systems provided on the Internet (some of them free), and
there will no doubt be more in the future. For the general public, computer software for translation is a quite new product; they are unaware of the advantages, limitations and methods of using such systems. They are furthermore familiar with rapid improvements of computer technology and software, and will therefore be expecting similar rapid improvements in the quality of translation software. One of the tasks of the MT community must therefore be to convey some idea of the immense complexities involved in dealing with natural language, and to explain why automatic translation has progressed so slowly despite over forty years of intensive research (Hutchins, 2000).

A software system is a “machine translation system”\(^{(3)}\) when input units are full sentences of one natural language and the output units are corresponding full sentences of another language. Here, a translation program should provide: a.) minimally correct morphology, b.) some minimal syntactic processing and c.) some semantic processing (ibid.).

Development of MT systems has continued in three main trends: a.) computer-based tools for translators, b.) operational MT systems involving human assistance in various ways and c.) pure theoretical research towards the improvement of MT methods in general. To reduce problems of ambiguity, researchers have proposed the development of controlled languages\(^{(4)}\) and the restriction of systems to specific subject areas. All current commercial and operational systems produce output which must be edited (revised) if it is to attain publishable quality. So, it is now widely accepted that MT proper works best in domain-specific and controlled environments (Hutchins, 2001 a: 2-3).

Thus, rather than adapting general-purpose MT systems, it is now recognized that it may be better in some circumstances to design systems for
use with controlled language. A number of independent companies have been doing this in recent years.

Hutchins (2001b) states that the principle criteria for assessing systems and aids are already and will be increasing in the future their a.) usefulness and b.) usability. The usefulness of a system or a tool relates to its basic functions and its aims. So, systems will be judged not (primarily) by the quality of their translation, but by their usefulness in achieving the aims they are designed for. Also, systems and translation aids will be assessed for their usability (e.g., ease of use) in the specific contexts and environments for which they are intended.

All in all, Hutchins (ibid.) confirms that “if we do not want MT prejudices and misunderstandings to continue, then it is in our hands to ensure that in the future our aims and objectives are well understood”. In his view, one way of doing this is to develop a new ‘vision’ of what is being done and why. He stresses his belief that most of the users (in particular, the general public) do not want automatic translators (or Fully Automatic High Quality Translation FAHQT systems), but aids for effective and appropriate cross-language communication.

2.11.2 MT Systems and Text Type Considerations

Hutchins (2001a:4) indicates that the principal focus of MT research remains the development of systems for translating written documents of scientific and technical nature. Outside the range of possibility are literary and legal texts, indeed any texts where style and presentation are important parts of the message.

It is generally stated that MT is suitable for a limited class of texts, variously called “non literary” or “informative”. In practice, this has meant
machine instruction manuals, technical articles, textbooks, minutes of meetings, etc. Texts suitable for MT are so because translation must involve computing only the information needed for defining equivalents. Thus, in texts which make reference to a real-world fragment common to SL and TL cultures, the quality of real-world knowledge that must be considered is minimized. In terms of word-level equivalents such texts may use a good deal of terms for which there are internationally standardized translations. The MT environment is ideal for ensuring consistency of such terminology. At the other end of the scale, stylistic equivalents are clearly defined by cross-linguistic conventions applicable to the particular text type resulting from the existence of international communities with common aims. Of course, the problems of word-level and supra-sentence-level equivalents remain to an extent sufficient to make MT for these text types far from trivial (Kilby in Whitelock (ed.), 1995 : 10).

Another aspect that makes informative texts particularly suitable for MT is the fundamental applicability of syntactic notions of well-formedness. If an MT system fails to provide an analysis for a sentence, it may be due to unclear rather than an incomplete grammar. If the MT system is embedded in an environment which provides interaction with the writer, it may be possible to elicit an alternative of greater clarity. In fact, in some environments, input to a system consists of texts drafted in a restricted language. Writers seem satisfied that this promotes genuine improvements in the quality of the SL text (ibid.).

In informative texts, intentional ambiguity does not play a significant part. Certain MT systems are intended to deal with a very limited, highly formalized text type, such as the weather reports handled by METEO\(^{(5)}\). In
cases such as these, the grammar–writing tools available to linguists may be unsuitable for more general text types (ibid.).

Finally, in systems which aim to translate texts in a variety of subject areas, the amount of knowledge incorporated will be vast. In such systems, the generality, flexibility and perspicuity of the grammar-writing tools is a paramount design aim, since the process of grammar writing consumes far more human resources than does the initial design of the system.

2.11.3 Classification of MT Systems

MT systems can be classified\(^6\) according to the text-type they can handle and the degree of automation. These two modes are briefly outlined here:

2.11.3.1 Typology of MT Systems by Text Type

By text type, two modes of distinction are common, one is based on the scope of the subject field, while the other on degrees of restriction of input. Regarding scope, a difference can be made between: (a.) systems based on a single fixed textual corpus (practical only for research purposes and very limited applications), (b.) systems limited to a narrow subject field (e.g., the case of METEO), (c.) systems of limited scope which envisage extendibility through the addition of new dictionaries and (d) unrestricted systems (Sager, 1994 : 289-90).

Regarding restrictions, a distinction is often made between systems that require to pre-edit input and systems which input texts written in a language designed for ease of application of MT (e.g., International Language for Service and Maintenance) (ibid.).

2.11.3.2 Typology of MT Systems by Degree of Automation
Sager (ibid:290) presents the following clarification of MT systems by degrees of automation:

a.) Machine-assisted human translation systems.

b.) Human-assisted machine translation systems.

c.) Fully automatic translation systems.

The first group is characterized by tools which assist the translators, namely:

(a.) Word processor + dictionary
   i. Selective dictionary look-up.
   ii. Automatic dictionary look-up.

(b.) Term-bank access.

(c.) KWIC indexes for the avoidance of homonyms.

(d.) Grammatical information.

(e.) Morphological analysis.

(f.) Access to translated – text field.

(g.) Automatic correction of spelling and typing errors.

The second group is characterized by the form and degree of human intervention:

(a.) Pre-editing on screen.

(b.) Interaction during translation.

(c.) Post-editing.

Finally, full automatic systems are defined as operating without human intervention between input and output.

2.11.4 Description of MT Systems

In this section, concentration will be on three English-into-Arabic MT systems that are being used and/or marketed today, i.e., AL-Mutarjim Al-Arabey version 3.00 (2002), Golden Al-Wafi version 1.00 (2002) and An-
Nakel Al-Arabi (known in the local market as AL-Kafi) version 2.00 (2000). These three bilingual translation systems are the subject of evaluation in this study. Other systems of the same kind are essentially the result of successful MT R & D (research and development) projects. It is necessary to give a minimal identification and brief description of these systems here. In fact, there is still a very limited number of MT systems that involve the Arabic language, but software as translation aids are more useful and widely used than these MT systems.

It is worth noticing that no serious evaluation tests and comparative evaluations of these systems have been performed since the actual and accurate evaluation processes should be carried out within the framework of the International Standards measurements/metrics, tools and scales. In addition, such evaluation processes require intensive work and efforts (mostly team work in workshops) that take long time where the system can be tested appropriately with respect to translations of various text types and of different levels of complexity (regarding both the SL & TL).

According to AL-Sieny (2000:177), comparative evaluations of MT systems should take into consideration certain aspects on which the tests will be based. Some of the most common aspects are:

1. Accuracy of translation.
2. Correctness of the target text (TT).
3. Ease of post-editing and modification of the translated text and ease of comparison with the ST.
4. Ease of entering additional modifications for the improvement of the system itself.
5. Types and sizes of the dictionaries in the system and the varieties of Arabic vocabularies to each English word taking into account all the
possible meanings and the possibility of selecting the most relevant and potential meaning.

(6) The time needed for the translation of various text types with respect to complexity and content of special domain.

(7) Cost of the system and usability, i.e., ease of use with respect to different types of computers and machines.

(8) Network communications to facilitate electronic publication and transfer.

2.11.4.1 Sakhr

It is the first Arabic company dealing with the use of Arabic in computer software and research programmes concerning the analysis of Arabic and MT. Sakhr software has provided a family of products that have played a major role in the field of MT starting with Sakhr CAT (Computer–Aided Translation) which is the translators’ assistant in huge translation tasks and Sakhr Enterprise Translation System (SET). The company has produced various kinds of software and translation aids such as: the Automatic Reader, Texts Systematic Organizer, Syntactic Analyzer, Computerized Arabic Dictionary, and recently the company has produced many systems for Internet translation and communication (e.g., Sindibad(1), Sindibad(2), Nashir Net, AL-Adreesi; AL-Dalil, etc. The company, in fact, has aimed at designing and producing Arabic MT systems that produce adequate and efficient output (AL-Shareef, 2000: 177).

In the design of MT systems, Sakhr follows the transfer approach\(^\text{7}\) in translation, i.e., a linguistic analysis of the ST, then transferring its components into Arabic and then generating the adequate Arabic text (TT). It utilizes various kinds of tools in building and improving its MT systems (e.g., syntactic analyser, automatic dictionary, text systematic organizer,
etc). The company has also designed interactive translation systems especially in the selection of appropriate lexical items, idioms and certain expression (ibid.).

2.11.4.2 Turjman

An MT system that makes use of a multilingual electronic dictionary (ADIB) based on Arabic–English and French. The system is produced by the Local Institute for Communication and Mass-Media Sciences in Tunisia. It has not been presented to the market yet, as it is still being improved by a group of experts (ibid.: 179).

2.11.4.3 ArabTrans

This system was developed by Arab Net Technology limited. According to its developers, ArabTrans translates texts from English-into-Arabic at more than a thousand words per minute, but the translation produced by the programme requires editing for both grammatical accuracy and alternative meanings preference (Arab Net, 1996, and Guessoum and Zantout, 2001: 136).

ArabTrans runs appropriately in Windows environment supported by Arabic. Texts can be entered into the system by: a.) a previously-prepared electronic file, b.) keyboard or c.) an electronic scanner of English texts. The operation of the system requires a memory of at least 4 Meg (better 8 Meg). It depends on the transfer approach in translation; although it practically relies heavily on lexical translation as it contains a large number of dictionaries (about 38) in addition to a special dictionary that can be added by the user in case of need. The general dictionary of this system includes (1,500,000) words. The system works with two main stages:
1.) a complete (full) translation of the whole text before post-editing; or
2.) partial translation of the text, i.e., paragraph-by-paragraph (a paragraph as a basic unit of translation).

The system has a ‘dictation-checker’ for the English text. Moreover, it enables the translator to post-edit the output and to search in the dictionary for the appropriate Arabic equivalents to each English lexical item. It is also possible to stop operating the system after the translation of each paragraph for post-editing and revision (AL-Shareef, 2000: 180).

In the manual of this system, the company mentions that it is important to say that this system cannot replace a human translator of literary texts, who can accurately transfer the spirit of the text. The translation it produces is still in need of certain modifications and corrections of linguistic errors and appropriate choice of lexical items (Arab Net, 1996).

In ArabTrans 2.0, the company has added new features and improved on some of the old ones. These include a new improved user interface, improved translation engine, smart toolbar icons, override meaning and a transliterate option, to name but a few. Texts can be entered from keyboard or disk. The Arabic translation is produced and displayed on the screen alongside the original English text. Translation can be achieved in seconds or minutes which allows the translator to double check the resulting translation and edit it to obtain the best results. ArabTrans second version supports text written in UK English and American English (Arab Net, 1998).

2.11.4.4 AL-Wafi

An MT system from English-into-Arabic only (i.e., uni-directional) developed by Arabic Trade Agreement (ATA) Software Technology Limited. It is a scaled-down version of AL-Mutarjim Al-Arabey (AL-
Shareef, 2000:181) and it uses the same MT modules as AL-Mutarjim AL-Arabey, except for a less extensive lexicon (Guessoum and Zantout, 2001). Thus, it has been excluded from the evaluation scheme of the present study especially after it has been found that it gives exactly the same results and percentages in the evaluation processes as AL-Mutarjim AL-Arabey system (AL-Bahadily, 2001, and Guessousm and Zantout, 2001).

In the User’s manual of the system (AL-Wafi 2.00, 1999), the company points out that Al-Wafi is designed for Arabic Windows to provide core English-into-Arabic text translation facilities, for those who do not need the advanced features of a product like AL-Mutarjim AL-Arabey. The system includes an English/Arabic translation dictionary containing over 200,000 English words, and over 750,000 Arabic words.

As AL-Wafi incorporates a fully-featured text editing environment, the English text can be entered directly into the system using the keyboard. The text can also be pasted from the clipboard or imported from a plain text file. The translated Arabic output can then be saved as a text file or edited within AL-Wafi and also copied and pasted into any Arabic-supporting word processor (AramediA Group, 2001).

To run AL-Wafi, the system requires a PC with a minimum hardware requirements of:
1.) A 486 DX 2 / 66 processor with 16 Mb of RAM.
2.) Arabic Windows 3.1 or Arabic Windows 95.
3.) An Arabic / English keyboard.
4.) A CD-ROM drive (at least double-speed).
5.) 20 Mb of free hard-disk space (AL-Wafi 2.00, 1999, and Aramedi A Group, 2100).
AL-Wafi incorporates a bilingual user interface, which can operate with menus and commands displayed in English or Arabic. The translation dictionary has over one million words in total, with the English dictionary based on the Oxford English Dictionary. The User Dictionary can be used to add the user’s own words and specialist terms to the main dictionary. It allows individual words to be translated from English-into-Arabic or vice versa. AL-Wafi can accept English text files of up to 40,000 characters, with longer documents being translated in sections. Its translation speed can reach over 1000 words per minute on a suitably specified system. It has a phonetic translation feature, which enables it to create Arabic equivalents of the names of people or places. It has fast and accurate English spell-checking facility, using its own English dictionary. With this system, translation results can be improved over the default translation, by selecting alternate words for a particular term. The system is not suitable at all for translations of literary texts. It can better translate business letters. The system fits more as an electronic dictionary than an MT system and it has been found that its major function is not translation but ‘Arabicization’ (Qendelft, 1997 and AramediA Group, 2001).

It is worth mentioning that ATA company has recently produced its latest versions of the English/Arabic software. These versions are supposed to offer great improvements over previous ones. Such new versions include: a.) New ‘Al-Wafi‘ 3.00 which is an English–Arabic translator designed to meet the demands of the ordinary user, b.) ‘Golden Al-Wafi‘ 1.00 a translation software for more advanced users (See 2.11.4.8.3) and c.) ‘Al-Mutarjim Al-Arabey’ 3.00 (See 2.11.4.8.1) for professional English-Arabic translations. Further, the company has also produced the first version of ‘Al-Wafi Quick Dictionary‘, which is an extensive English/Arabic–
Arabic/English dictionary designed to be a companion to those who need constant dictionary look-ups whilst maintaining the flow of the rest of their work (ATA Software Technology, 2002).

2.11.4.5 SYSTRAN
SYSTRAN Company has started its multilingual MT systems since the beginning of the 1980s, supported by the American Pentagon. It has produced a large number of language-pair MT systems that can run on “Manie Frame”, i.e., large computers. Some of these systems deal with Arabic (AL-Shareef, 2000: 183).

In the mid 1990s, the company has transferred the previously-developed systems to be used on PCs after improving them. Their major concern has been in the widely used languages in the European Market (e.g., French, English and German). The company has not yet undertaken the production of Arabic MT systems, although this is one of the main plans of the company. The system that can run on Maine Frame was tested. The results show that it is a primitive system of poor output and it needs great efforts not only to make it run on PCs, but to improve its capabilities in Arabic as well (ibid.).

2.11.4.6 Applied Technology (Apptek)
This company has started to develop MT systems and translation aids and tools since 1990s cooperating with a group of experts in AI and Arabic in Amman/Jordan and Rabat/Morocco. It has developed MT systems that depend on: a.) the linguistic analysis of the ST and, b.) transfer into the TT. The whole process begins by a detailed linguistic analysis of the ST (a morphological and syntactic analysis) with the aim of dividing the text into
its main linguistic components: words, phrases, clauses/sentences on the basis of the modern theories of CL especially lexical/functional grammar. Then, the system transfers these components (after a complete comprehension of their meanings) into the TL (Arabic). Next, it generates the first translated draft of the original text. After this stage, the system starts to examine carefully the language for correction of errors and removal of ambiguity. It is worth mentioning that the system, when generating the Arabic text, takes into account the user’s environment culturally, socially and technically. The company aims at producing full MT systems. It has produced sample scaled-down versions of these systems which can operate on PCs, first displayed in Gitex Fair in United Arab Emirates (UAE), 1997 (ibid: 183-84).

The system includes a general dictionary of more than 100,000 words. It also includes special domain dictionaries of business, law, medicine, technical science, military, etc. The user can build his own dictionary too. There is a dictionary of idiomatic expressions in this system and for the sentences frequently repeated in translation. This is to enrich the translation memory of the system. There are other translation aids and tools in this system. The company is seeking to develop this system for on-line translation of texts and other relevant applications. It believes that the best MT system can achieve 50%-70% accuracy in translation depending on the level of complexity of the ST, the specific domain it deals with, and the cultural and social environment where the ST and the TT should coincide (ibid.).

Further, the company is about to produce an on-line Arabic scanner cooperating with other companies (e.g., Microsoft).
2.11.4.7 Other Arabic MT Systems and Projects

In addition to the above-mentioned MT systems, there are other special-domain MT systems and projects such as:

2.11.4.7.1 Electronic Research Institute: The National Research Council in Cairo

This institute designs special-domain MT systems in cooperation with the European Group for Translating Medical Texts (ARAMED) using the transfer approach in translation. It is restricted to the medical context. It is part of a project with the German (IAI) Institute and the French (Talana) Institute for MT known as (CAT₂). The institute in Cairo is responsible for the Arabic programme of the project which is currently restricted to MT into Arabic. The institute aims at developing MT systems that go from Arabic into different European languages in the future. The project involves building a lexicon 60% specialized in medical terms and 40% in general terms. In addition, the institute is working with another project of building a multilingual based on corpus, as well as with a project of identifying Arabic letters and generating special Arabic fonts (ibid: 185-86).

It is worth mentioning that the Translation Institute in AL-Azhar University has recently started a number of MT projects and development of MT tools and aids as a partial requirement for a higher degree in the Institute.

2.11.4.7.2 The Scientific Center in Cairo (IBM)

The Scientific Center in Cairo under IBM has produced a special MT system known as “Translation Manager” used in IBM only (it has not been
marketed yet). It is still in the stage of development. There is another system produced by IBM too, to identify and recognize Arabic speech-sounds and other online Arabic systems (ibid : 186).

2.11.4.7.3 XEROX Company
The company has produced an analyzer of Arabic morphology and sent a testable version to the market. The system with the number (501) takes Arabic words and analyzes them morphologically. It gives results of different forms for each single word taking into consideration its derivations, inflections and different meanings in English (ibid.).

2.11.4.7.4 Luts Company
Luts Company has recently produced an Arabic version of “Luts” for on-line services.

2.11.4.7.5 The House of Computerizing the Arabic Text
This company in Jordan has produced an MT system known as the “Arabic-light Pen Program” to identify the printed or manual Arabic text (ibid.).

2.11.4.7.6 Microsoft
The company has a large project to produce Arabic versions of all its translation systems and aids (ibid.).

2.11.4.8 Arabic MT Systems Under Evaluation
Only three English-into-Arabic MT systems that the researcher of the present study has managed to purchase and obtain are under evaluation. These are a.) Golden Al-Wafi 1.00 (2002), b.) AL-Mutarjim AL-Arabey
3.00 (2002) and c.) An-Nakel AL-Arabi 2.00 (2000). A detailed description of each is presented below.

2.11.4.8.1 AL-Mutarjim AL-Arabey

2.11.4.8.1.1 General

A uni-directional English–into-Arabic MT system. It was first produced by ATA Software Technology Limited in 1996 and the second revised version of the system was produced in 1997. A scaled-down version of this system is known as AL-Wafi (also produced in 1997) (Al-Shreef, 2000 : 180).

The company claims that it is “the first English-Arabic MT system ever to be developed on personal computers” (ATA Software Technology, 1997). Version 2.00 of AL-Mutarjim AL-Arabey is a “giant leap forward in Arabic software translation. All kinds of documents, articles, reports, or books can be translated easily using the powerful features on offer” (ATA Software Technology, 1999).

The first version of the system has a general dictionary of 300,000 words, which has been expanded into 2,000,000 words in the second version to enable the user to get the best possible quality of translation of documents from English-into-Arabic. The system can translate multi-document text regardless of the length. A minimum translation speed of the system is 1000 words per minute achieved with just the basic hardware requirements. The electronic dictionary of the system goes from English-into-Arabic and vice versa. A scanning and Optical Character Recognition (OCR) software is incorporated, so that existing paper-based English texts can also be easily translated. The original file formatting (tabs, spaces, brackets, etc) of the English text is maintained in the translated Arabic text. The English text can be spell-checked for accuracy before translation. A special user dictionary can be added for defining new words or special terms. There is, also, a
temporary dictionary for the current document. The system contains
dictionary of abbreviations for translation of abbreviations such as BBC,
MBC, etc. It can load up six dictionaries at any one time and bilingual
interface with commands and online help in both English and Arabic. A full
text editor is included for entering, importing or changing the English or
Arabic text without having to use a separate word-processor. With AL-
Mutarjim AL-Arabey, selected (highlighted) text can be translated on its
own without affecting the rest of the already translated/edited Arabic text. In
this system, there is a synchronous scrolling between English and Arabic
text so that the same parts of both documents can be compared easily.
Moreover, specialized dictionaries are available for business, commerce,
finance, military, social science and water engineering topics (ibid. and AL-

According to AL-Shareef (ibid.), the ST (input) can be entered into the
system by one of the three common ways: a.) a ready electronic file, b.) the
keyboard or c.) an electronic scanner. The system also has a dictation-
checker for the English text. The text is fully translated by the system before
post-editing and revision. A partial translation, i.e., paragraph -by- paragraph
is also possible where the user can revise and post-edit each paragraph once
it is completely translated by the system. The translation with this system is
performed in two stages: a.) the system identifies the English text (ST) then
b.) it starts parsing and analyzing it. On the basis of this, it begins to
formulate the Arabic text (TT) according to the Arabic morphological and
syntactic patterns for an adequate translation.
This system permits some colloquial English expressions and proverbs in the ST, but it is based mainly on Standard British English rejecting, for instance, American English and other varieties of English (ibid.).

To be able to run AL-Mutarjim AL-Arabey 2.00, it needs a personal computer (PC) with a minimum hardware requirements of:

a.) A 486 DX2 / 66 processor with 16 Mb of RAM.
b.) Arabic Windows 3.1 or Arabic Windows 95.
c.) An Arabic / English keyboard.
d.) A CD-ROM drive (at least double-speed).
Eidan and Ibrahim (2000, 289-90) maintain that the English texts (STs) that the system accepts must be of the type (.TXT) coming from different sources as seen in the diagram above.

For the words that are not found in the general dictionary, the system can check them by using the English dictation-checker. If it fails to find them, it leaves them as they are in the Arabic TT. The system relies in its operation on the following components:

a.) English – Arabic dictionary.
b.) English syntactic analyzer.
c.) Arabic morphological analyzer.
d.) Arabic syntactic analyzer.

2.11.4.8.1.2 The General Dictionary

The general English/Arabic dictionary of the system is designed to deal with single English words in addition to terminologies that are composed of more than one word. It provides all the necessary information AL-Mutarjim AL-Arabey requires in the translation process. It gives the stressed and unstressed-Arabic roots of the English verbs. For instance, the verb “كتب” (=write) is unstressed; while “شدّد” (= stressed) is stressed (ibid : 292).

2.11.4.8.1.3 How Does AL-Mtarjim AL-Arabey Work?

The translation process in AL-Mutarjim AL-Arabey involves first of all the analysis of the English sentence, with the help of the English syntactic
analyzer and the dictionary, into noun, adj., verb, etc. and the syntactic functions of these components in terms of subject, verb, object, etc. The position of each component in the sentence is also indicated (Fig.2.1). The Arabic morphological analyzer gives the conjugation of the Arabic words; while the Arabic syntactic analyzer determines the syntactic functions of the Arabic word and their positions in the sentence. In addition, it vowelizes these words correctly. Then, the system displays the Arabic translation putting the words in their correct syntactic position in the sentence matching the Arabic word order and other syntactic aspects such as the occurrence of the verb before the subject, the noun before its attribute, etc (ibid : 292-93).

Now, if we take the English sentence ‘The boy laughed’ (Fig.2.2), for instance:

![Tree-Structure of the Sentence](image)

**Figure (2.2). The Tree-Structure of the Sentence.**

The English syntactic analyzer with the English-Arabic dictionary of the system starts analyzing the given sentence showing its basic components. The sentence consists of two phrases:

1. a noun phrase that consists of the definite article (the) and the noun (boy);
2. and a verb phrase represented by the intransitive verb (laugh).
Syntactically speaking, it is a sentence as it is composed by the subject ‘boy’ and the verb ‘laugh’. In the lexicon, the noun ‘boy’ which is a single, masculine, animate noun is found in the third person singular form. The verb ‘laughed’ is a simple past tense verb.

The information available is transferred to the Arabic morphological analyzer which gives the correct conjugation of the Arabic word according to the root given in the dictionary. Then, the Arabic syntactic analyzer determines the analytical forms of the words in the sentence and their appropriate vowels. After that, it puts these forms in the correct order that suits the Arabic sentence. On the basis of this, the reading of the sentence translation of the given English sentence, as:

\[
\text{ضَحِكَ الولدُ}
\]
The following figure (2.4) shows how the system runs when translating the sentence:

Figure (2.4). The Whole Translation Process in AL-Mutarjim AL-Arabey When the Verb is Intransitive / Simple Past Tense.

<table>
<thead>
<tr>
<th>English Text</th>
<th>The boy laughed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic Text</td>
<td>ضحك الولد</td>
</tr>
</tbody>
</table>

**General Issues:**

AL-Mutarjim AL-Arabey deals with many general issues related to English texts and the requirements of their translation into Arabic such as (ibid:300-301):
1. Hyphenated Expressions: The system can distinguish between words that can be either hyphenated or non-hyphenated such as the adjective ‘first-class/first class’ and the adjective ‘hand-held/handheld, and these that must be hyphenated (e.g., the noun ‘A-bomb’ referring to ‘atomic bomb’ to be distinguished from ‘bomb’ in general in ‘a bomb’).

2. The position of the Definite article ‘ال’: This concerns the Arabic expressions of more than one word. So, the expression whose second word is only to be identified (e.g., ‘غرفة نوم’/’bedroom’) is correctly translated into ‘غرفة النوم’ if preceded by the definite article ‘the’. The same thing happens with the expression whose both words are to be identified (e.g., ‘القيمة المطلقة’/‘the absolute value’), although these two words are not originally definite in the dictionary. Accordingly, the dictionary entry ‘غرفة نوم’ is correctly translated in the examples below:

   e.g.,
   (1) The cat is not in the bedroom.  القطة ليست في غرفة النوم.
   (2) The third bedroom is cleaner.  أن غرفة النوم الثالثة أنظف.
   (3) This is my new bedroom.  هذه غرفة نومي الجديدة.

3. Prefixes and Suffixes: AL-Mutarjim AL-Arabey deals very well with prefixes and suffixes generating an infinite number of correctly translated words even if they are not found in the dictionary. There is no entry in the dictionary for “pollution-free”, for instance; but the system grammatically processes this expression as an adjective and translates it correctly, as in:

   e.g.,
(4) We live in Pollution-free town.
(5) My brother is an ultra-ambitious young man.

4. Transliteration: The system processes proper nouns that are not found in the dictionary and transliterates them including the Arabic names with accents that correctly appear in the translation, as in:

\[ \text{e.g.,} \]

\[ \text{‘Aqeel} \quad \text{أقبل} ; \quad \text{As Salhi} \quad \text{صالحي} \]

\[ \text{Sa‘ad} \quad \text{سعد} ; \quad \text{Tariq} \quad \text{طارق} \]

5. Dictionaries: In addition to the general dictionary, AL-Mutarjim AL-Arabey uses many dictionaries that help in giving the most potential Arabic equivalent to the English expressions and terminologies. These include:


b. Current Documents Dictionary: is only restricted to the terminologies relevant to the document being currently translated.

c. Dictionary of Abbreviations: is restricted to terminologies written as abbreviations and acronyms, e.g., LO and OPEC.

d. Restricted/special-Domain Dictionary: is restricted to terminologies of special domains on arts and science, etc.

2.11.4.8.1.5 Al-Mutarjim Al-Arabey 3.00 (2002)
ATA has made certain improvements on and additions to version 2.00 of Al-Mutarjim Al-Arabey that appear in the third version of the system. It is designed as a professional MT tool for more advanced users, that can
quickly and easily translate a text from English-into-Arabic. Al-Mutarjim Al-Arabey 3.00 is ideally suited to translating long documents and thousands of words in only a few seconds (a minimum translation speed of 3000-10000 words per minute) with just the basic hardware requirements (PC, 166MHz, 32 or 64 Mb RAM, 50 Mb free HD space, optional Sound Card, Arabic Windows 98/2000/ ME/NT/XP). With Al-Mutarjim Al-Arabey 3.00, users will no longer have to do hours of tedious looking through dictionaries. All kinds of documents, articles reports or books can be translated easily using the powerful features on offer. These features are listed below (ATA Software Technology, 2003 a):

1. The system has a new and advanced translation engine.
2. It can read the English text (English Text – to – Speech).
3. It has a new dictionary with more than two million English and Arabic words (General Dictionary).
4. It contains forty-six specialized science dictionaries in medicine, veterinary, biology, law, military, mathematics, engineering, ecology, etc. In addition, thousands of scientific terminologies are included under ‘other science‘. These dictionaries are meant to help the user (student, researcher, etc) as references only and are not used by Al-Mutarjim Al-Arabey in the translation of texts. Words and terms used in these dictionaries have special Arabic translation which could be totally different if they were translated using the general dictionary. So, it is highly recommended that professional users should use a special-subject dictionary for more accurate translation.
5. The system can translate multi-document text regardless of the length.
6. It can translate a number of long documents in the background (batch/direct translation), where the English text and its Arabic translation appear in one document.

7. Al-Mutarjim Al-Arabey has an English spelling checker and the original file formatting (tabs, spaces, brackets, etc.) of an English text is maintained in the translated Arabic text.

8. The system can send the translated text by e-mail. Also, it can translate selected texts from the Internet or other programmes in the background. Files received by modem or fax cards can be stored as text files and then translated.

9. Al-Mutarjim Al-Arabey has a powerful built-in scan-OCR software for scanning English documents and converting them into text files ready for translation.

10. It can save the translated files in one output file. This will make it easy to print all files at once, for example.

11. The system can translate English texts that have the extension doc/wri/ps/rtf/ in addition to (TXT.) form as an ASCII file.

12. Al-Mutarjim Al-Arabey 3.00 has a bilingual user interface and a bilingual user guide. It also has a bilingual dictionary English/Arabic and Arabic/English.

13. To get the best and most accurate Arabic translation, Al-Mutarjim Al-Arabey 3.00 can use different dictionaries that are created by the user:
   a. Preference Dictionary: where the user can create a preference dictionary of preferred Arabic meanings for the English entries in the general dictionary. These preferred meanings can be added to the dictionary while translating a certain document or text. For example, the word ‘development’ has four Arabic meanings listed as follows:
The user can build a list of these preferred words with respect to the theme of translation to help enhancing the quality of the translation. This dictionary can be used whenever the user intends to translate a text of the same theme, as the system will select the preferred meanings automatically.

b. The User Dictionary: users can create their own dictionaries. They can be used to enter special words or terms which will be used in the translation overriding the words in the general dictionary.

c. Abbreviation Dictionary: This dictionary helps the user to keep a list of abbreviations and their complete wording to be used by Al-Mutarjim Al-Arabe in the translation of texts.

14. This version of the system consists of a list of themes that the user can select before translating a text, so that the system can automatically choose the Arabic equivalent to the English word with respect to the subject matter of the text. The word ‘fan‘, for instance, has the Arabic meanings (مشجع, مؤيد, معجب, نصير) where the second word ‘مشجع‘ fits in a sport context and the last meaning ‘معجب‘ in a musical context. In translation, Al-Mutarjim Al-Arabe selects the meaning that goes with the theme of the text.

15. As in any MT software, and indeed in human translation text, post-editing the TT is a must. Al-Mutarjim Al-Arabe includes a routine for displaying alternative Arabic meanings to the same English word to enable the user to select the correct Arabic word.

16. It is worth noticing that Al-Mutarjim Al-Arabe 3.00 can be registered on one computer only, i.e., it cannot be re-registered on another
computer. But, a user can re-register it on the same computer for up to ten times in case something goes wrong with the computer or re-installing Windows.

17. Al-Mutarjim Al-Arabey 3.00 is protected against piracy and illegal copy and distribution by the requirement to register a copy of the system with the developer and copyright holder, ATA Software Technology Ltd.- through their website at: http://www.atasoft.com. Each owner of the system is given a certain serial number and the installation and registration of the system requires Internet connection. Such connection to the Internet is not needed to run the system.

18. Finally, this version of the system can be freely updated through the website of ATA company.

![Figure (2.5). The Main Window (and Standard Toolbar) of Al-Mutarjim Al-Arabey 3.00.](image)
Figure (2.6). The Setting Window of Al-Mutajim Al-Arabe\textregistered~3.00.

Figure (2.7). Translation of Proper Nouns and Abbreviations.

Figure (2.8). Diacritics on the Arabic Words.
Figure (2.9). Alternative Words in Al-Mutarjim Al-Arabey 3.00.

Figure (2.10). Translation of Words According to the Theme of the Text.
2.11.4.8.2 An-Nakel AL-Arabi

2.11.4.8.2.1 General

An-Nakel AL-Arabi is an MT system produced by Cimos Company for software technology in France. Since 1980s the company has developed MT tools and systems that deal with Arabic. AL-Mutarjim AL-Kafi is the first system produced by the company translating from English-into-Arabic and the reverse. As a representative of Cimos (2002) indicates, this system is a very old obsolete version of An-Nakel. The company has developed another system known as MLTS (Multilingual Translation System) which is used for on-line translation.

An-Nakel is a bi-directional MT system for Arabic whose approach is completely different from the word-by-word translation commonly found in electronic translation. It works with the sentence as the basic unit of translation in a rule-based processing system and knowledge database, with a strong parser, deep syntactic analysis and selective semantic analysis to detect main verb, phrasal verbs and idioms. It works in both batch (stand
alone translation programme) and interactive mode (tool to help translator), with menus in English or Arabic, and is capable of translating 60,000 words per hour depending on the processor’s speed. It can translate a complete file, or just a paragraph or sentence, transliterates proper nouns, handles abbreviations and manages multiple meanings and has a dictionary of 150,000 entries that can be extended by the user (An-Nakel, 2001 and Renouf, 2001).

This system is a sophisticated state-of-the-art. It requires PC with 486, Pentium or greater; MS Windows 2000, Windows NT 4.0 (SP3 or later), Windows 95/98 or later (Arabic version required); 16 MB RAM; 100 MB hard disk space for typical installation; 64 MB available for swap space; graphics card; colour monitor (An-Nakel, 2001 and ITP Net, 2002).

According to An-Nakel (2001), the system is a computer-assisted–translation software designed as a tool to help translators and not to replace a skilled human translation. An-Nakel AL-Arabi is precise and fast for producing a draft translation of a high quality. It can be used for electronic translation of technical documentation and manuals. It also produces computer-aided translation in a wide range of subject areas like science and technology, commerce and banking, computer and petroleum. Since An-Nakel AL-Arabi translates in a specific subject area with customized dictionary and translation-memory database, the TT is acceptable and understandable with a good accuracy. Users can edit TTs using the standard word processors (Microsoft Word). Because of its deep sentence analysis and semantic connections, An-Nakel AL-Arabi is able to learn new rules and knowledge. In other words, it learns new rules when a translation for a particular proposition or sentence is suggested, and is able to use them when
a similar structure occurs. This greatly improves the quality of the translation.

2.11.4.8.2.2 How It Works

An-Nakel AL-Arabi has been developed by an experienced team of translators, linguists and engineers. It combines new concepts in multilingual processing, MT and Translation Memory (TM). Since the sentence is the basic unit for translation process, each is analyzed deeply. An-Nakel AL-Arabi determines the different syntagms and categories of words. It attempts to analyze words in context and then builds semantic links. The analysis process is ended by an internal representation of the sentence. The TT is generated by the transfer method according to the grammar rules of the TL. In the case of translation from English-into-Arabic, the generated Arabic text can be displayed (printed) with or without vowels (An-Nakel, 2001).

An-Nakel AL-Arabi approach is fully different from word-by-word translation. The translation process is driven by deep sentence analysis and semantic meanings of words (locutions) used in a specific subject area. The system uses five dictionaries:

1. The Translation Memory Database: contains sentences already translated.
2. The General Words Dictionary: contains words in common use with approximately 100,000 words and locutions.
4. The Specific Dictionary: contains words used by specialists in a selected subject area.
5. The User Dictionary: contains words added or inserted by the user.

It is worth noticing that An-Nakel AL-Arabi dictionary covers the following subject areas:
1. Trade and commerce.
2. Banking and law.
3. Information and computers.

In this MT system, the following pairs of languages are available:
1. English into/from Arabic
2. English into/from French
3. French into/from Arabic

The user can enter and code his own specialized dictionary. The good points of the system are: speed and accuracy, while the bad point is the need for post-editing of the output. However, it is a much needed application and a gem to use (ibid. and *ITP Net* 2002).

An-Nakel AL-Arabi offers two operating modes:

a. On-line or
b. Batch mode.

In the first, the user may input and translate a document sentence by sentence. Here, the user can change the syntactic analysis and select the appropriate meaning of words with multi-meaning. The result of the translation is more accurate. In the second mode, the user submits the entire source document and the TT is stored in a separate file. In this case, the user can also update the TT using the “Alternative Meaning” in order to select the right meaning of words with multi-meaning (*An-Nakel*, 2001).

### 2.11.4.8.2.3 Levels of System Operation

As Azzeldine (2000:319-26) maintains, An-Nakel AL-Arabi operates on the following six levels:
1. TM database.
2. The level of morphological analysis.
3. The level of syntactic analysis.
4. The level of semantic analysis.
5. The level of transfer.
6. The level of synthesis / generation.

In the translation process, the system combines the TM database and the transfer approach. It is the only system that has such a property. The translator can also enrich the TM and the system builds the semantic relations and the deep structure (internal representation of the sentence).

The following is a detailed identification and explanation of the function of each of these levels:

1) **Translation Memory Database**:

Here, the system starts searching in the TM database for a sentence that corresponds the sentence under translation. It performs this by fuzzy matching, i.e., approximating between two series of letters. When the percentage of approximation reaches 100% or less, this means that the two series are correspondent and similar. If the percentage is not so, this indicates that the sentence being translated has a simple difference on the level of nouns, pronouns, verbs, conjugation, etc.

Then, the TM presents approximate suggestions by words (e.g., figures and proper nouns) or words that vary in their inflections.

The output of MT can be considered a complete group for feeding the TM database. However, this approach causes rapid saturation of the TM and increases the period of searching which requires finding a new style of ordering, arranging and controlling data. For translating the following Arabic sentence, for instance:
e.g.,

The man is writing a letter

the necessary substitutes are distinguished appropriately and on the basis of
the context. These could be:

e.g.,

1. The girl writes the letter.

2. The famous journalist writes the article.

3. The man has a book.

4. The man has a good book.

As the words in the two phrases “the famous journalist” and “the girl” have
the same classification (i.e., NPs) and the same can be said about “the letter”
and the article”; the two sentences are considered correspondents on the
level of TM. This leads to the storage of one of them only.

2) The Level of Morphological Analysis

An-Nakel AL-Arabi analyses each word in the sentence separately to
identify its morphological characteristics with respect to prefixes, suffixes,
number, gender, etc. Then, the system classifies words as:

1. Nouns according to the singular form.

2. Verbs according to the bare infinitive form.

3. Adjectives according to the singular masculine form.

4. Idiomatic expressions according to the base form of their verbs and the
third person singular of their pronouns, etc.

After that, the system searches for the longest series in comparison with the
stored elements in the following group of dictionaries:

a. Dictionary of general vocabularies and common expressions.
b. Dictionary of idiomatic expressions.
c. The user’s dictionary restricted to the vocabularies modified and added by the user of the system.

Each element in the sentence is distinguished by the linguistic classification it represents, i.e., as a simple word fragment, or idiomatic expression.

3) The Level of Syntactic Analysis (Parsing)

An-Nakel AL-Arabi analyses the structure of the sentence to distinguish the syntactic functions of its words and show how they fit the syntactic rules it contains. It, then, determines the function and nature of each element in the sentence and its division into grammatical units. After that, the system recognizes the syntactic functions in accordance with their semantic roles drawing a tree-structure of the sentence.

The process of analysis involves the identification of the verb and all the elements it requires, as the verb is the most essential element in the sentence; while the other elements (e.g., subjects, objects and others) are variables to the significant verb element. The other elements are distributed in the sentence according to the requirement of the verb and the syntactic functions are specified and determined accordingly.

The parsing process begins with the distinguished elements (variables) on the basis of their classification as NP, AP, VP, PP, etc or particles as (إنَّ،كان،لا…). So the sentence:

أطلب العلم من المهد إلى اللحد

is analyzed as:

```
[diagram]
```
Another problem concerns the relation between sentences with respect to meaning. Pronouns, for instance, should match, with respect to ‘gender’ and ‘number’, the noun they replace in the preceding or subsequent sentences to the sentence being analyzed. Thus, in the following sentence:

قدمت زينب هدية قيمة لصديقتها ففرحبت بها كثيراً.

the pronoun ‘ها’ in the word ‘هدية’ / present’ that occurs at the beginning of the sentence. The ambiguity results from the syntactic structure.

4) The Level of Semantic Analysis
Most of the MT systems available are concerned with the morphological and syntactic analysis of the Arabic sentence and there are still no systems that can handle the processing of the semantic analysis.

According to Azzeldine (ibid:321), An-Nakel AL-Arabi is designed to be the first Arabic MT system that takes this aspect into consideration.
Semantic analysis is one of the most important stages of the computers recognition of natural languages especially Arabic. It is a process of finding a semantic representation of sentences. The sentence is a combination of words that express a complete idea. It consists of either one main clause or a main clause and subordinate claus(es), i.e., a complex sentence or it contains two coordinated clauses (i.e., a compound sentence). The aim of such an analysis is to determine the meaning of each word in the sentence with respect to the context, supported by lexical and semantic information and dictionaries of idiomatic expressions.

Semantic analysis relies on a deep analysis of sentences to obtain the correct senses of words on the basis of their subject/predicate relations. There is a kind of recognition of the syntactic functions according to their semantic roles as subject, predicate and attribute. In addition, each word may have multiple senses that should be distinguished with respect to context. The word 'أعطى', for instance, or 'سلام' has different senses on the level of parts of speech and meaning depending on the context of the sentence. The verb 'سلام' has multiple senses that vary whenever the particle (a preposition) accompanies the verb changes. It has the senses: 'salute someone' / سلّم على ; 'admit something' / سلّم بين ; 'commit ones cause to' / سلّم أمره إلى . etc. The verb 'أعطى', on the other hand, also has different senses with respect to the word that follows it, as in: 'give something to someone' / أعطي فلانا شيئًا ; 'give lessons' / أعطي دروساً ; 'allow to speak' / أعني له الكلمة ; 'set an example' / أعطي مثلًا , etc.
In such an analysis, the linguistic rules and the sentences that have one sense or multiple senses are taken into consideration.

As with the idiomatic expressions, like:

القمر في كبد السماء، or وقع في حيص بيض، these cause two main problems:

a. The limits of the expression (i.e., its beginning and end); although it has many words, it is treated as one lexical unit and its parts should not be separated.

b. The semantic aspect of the expression, i.e., it is not easy to know if the intended meaning is the literal or the idiomatic.

An-Nakel AL-Arabi determines the beginning and end of these expressions and selects the correct idiomatic meaning.

In all the above-mentioned cases, the translator, in addition to the semantic selection that An-Nakel AL-Arabi presents, can add or modify another sense with respect to the context of the text. This is called “interactive” translation. In a sentence like ‘He is generous’, for example, one cannot tell whether ‘انه كريم’/ ‘He is generous’, for example, one cannot tell whether ‘كريم’/ generous is a proper noun or an adjective in Arabic. So, a human translator should interfere here to remove ambiguity.

5) The Level of Transfer
The stage of transfer in An-Nakel AL-Arabi consists of two steps:

a. Lexical transfer.

b. Structural transfer

The structure of the sentence in the ST is represented by a tree-diagram, and each tree-diagram in the SL has its counterpart in the TL. This process is important especially in English where the sentence should start with a
subject to be followed by a verb and an object (or complement) as in the example below:

1. The artist drew the painting  
رسام الفنان اللوحة.

Verb ‘to be’ in English goes with the noun phrase when translated into Arabic as in:

2. The man is always thirsty for knowledge.  
الرجل دائماً متعطش للمعرفة.

6) The Level of Synthesis (Generation)

The level of synthesis is different from that of transfer, as all the processes of morphological, syntactic and semantic analysis take place in the opposite direction (See Fig. 2.12 below). Here, the syntactic generation depends on the syntactic and semantic relations, where sentences are syntactically formulated with respect to placing words/phrases and expressions in their appropriate positions according to the linguistic rules of a language. The morphological generation, involves formulating words in their final correct shape such as assigning verbs the appropriate inflections of tense and agreement with subject (number and gender, etc).

<table>
<thead>
<tr>
<th>STEP 1</th>
<th>قاعدة التصوص المترجمة</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Translation Memory</td>
</tr>
<tr>
<td>STEP 2</td>
<td>نص المصدر</td>
</tr>
<tr>
<td></td>
<td>Source Text</td>
</tr>
<tr>
<td>ANALYSIS</td>
<td>Morphological Analysis</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------</td>
</tr>
<tr>
<td></td>
<td>التحليل الصرفي</td>
</tr>
<tr>
<td></td>
<td>Syntactic Analysis</td>
</tr>
<tr>
<td></td>
<td>التحليل النحوي</td>
</tr>
<tr>
<td></td>
<td>Semantic Analysis</td>
</tr>
<tr>
<td></td>
<td>التحليل الدلالي</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STEP 3</th>
<th>التحويل</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Transfer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SYNTHESIS</th>
<th>Syntactic Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>التوليد النحوي</td>
</tr>
<tr>
<td></td>
<td>Semantic Generation</td>
</tr>
<tr>
<td></td>
<td>التوليد الدلالي</td>
</tr>
<tr>
<td></td>
<td>Morphological Generation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>STEP 4</th>
<th>النص الهدف</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Target Text</td>
</tr>
</tbody>
</table>

Figure (2.12). A General Overview of An-Nakel AL-Arabi.
Figure (2.13). The Main Window (and Standard Toolbar) of An–Nakel Al-Arabi 2.00.

Figure (2.14). Document Commands of An-Nakel Al-Arabi.
Figure (2.15). Translation Menu Commands.

Figure (2.16). Dictionary Menu Commands.

Figure (2.17). Option Menu Commands.
Figure (2.18). Alternative Meanings for Words and Idioms in An-Nakel Al-Arabi.

Figure (2.19). Interactive Translation in An-Nakel Al-Arabi.

2.11.4.8.3 Golden Al-Wafi 1.00 (2002)
Golden Al-Wafi is the second Arabic MT System produced by ATA Software Company in 2002. Like Al-Mutarjim Al-Arabey 3.00, it translates English texts into Arabic only and not the other way around. These texts should be ASCII format (.TXT). This system can be registered and updated in the same way as Al-Mutarjim Al-Arabey 3.00. It is also protected similarly (See, 2.11.4.8.1.5). It is worth noticing that Golden Al-Wafi is not a new version of ‘Al-Wafi‘ MT system (ATA has also produced the third version of this system ‘New Al-Wafi 3.00 in 2002). Golden Al-Wafi is used for more advanced purposes than ‘New Al-Wafi‘, and it has the following general specifications (ATA Software Technology, 2003 b.):

1. Golden Al-Wafi runs under Arabic Windows 98/200/ME/NT/XP. It has a bilingual interface.

2. It has a new translation engine and the minimum system requirements are:
   i. 32Mb RAM - 64 Mb preferable .
   ii. 30 Mb HD.
   iii. CD-ROM Drive .
   iv. SVGA Colour Monitor .
   v. PC with 133 MHz .

3. The system can translate full texts, parts of a text (partial translation) and it can give the meanings of the highlighted words .

4. It can translate a number of long documents in the background (Batch translation).

5. Golden Al-Wafi displays Arabic texts with or without diacritics, and it has an English spelling checker, text alignment and it can read the English text .

6. This system has the following dictionaries :
i. General Dictionary: with more than two million English and Arabic words.

ii. Bilingual Dictionary: which can be used for special purposes by the user whenever he needs to know the meaning of an English word in Arabic or vice versa.

iii. Specialized Dictionaries: it contains eight specialized dictionaries in Medicine, Veterinary, Biology, Physics, Mathematics, Chemistry, Engineering and Geology, which are not used by the system in the translation of texts. They are intended for the benefit of the user merely.

7. Golden Al-Wafi can translate thousands of words in seconds with a minimum translation speed of 3000 words per minute.

8. It can perform multi-document translation.

9. The system comes with built-in routines for grammar checking in both languages, English and Arabic. It starts by checking the English text grammatically and parsing the sentence. Then, it passes the results to the Arabic parser to build the Arabic translation, taking into account the structure of the Arabic sentence before displaying the Arabic text.

10. Golden Al-Wafi has the capacity of transliterating proper nouns and names of towns, people, medicines, etc. In addition, it can translate abbreviations and transliterate them too.

11. It includes a routine for displaying alternative Arabic meanings to the same English word to enable the user to select the correct Arabic word. This is usually done after the translation of the text.
Figure (2.20). Main Window (and Standard Toolbar) of Golden Al-Wafi.

Figure (2.21). The Setting Window of Golden Al-Wafi.
Figure (2.22). Suggestions for Spelling Checker.

Figure (2.23). Multiple Meanings of Word.
Figure (2.24). Alternative Meanings of Word.

Figure (2.25). Diacritics on Arabic Words.

Figure (2.26). Translation and Transliteration of Proper Nouns and Abbreviations.
2.12 The State of Arabic MT

The widespread use of computer networks and the impact this has had on the quality of information production is putting insurmountable pressure on nations whose first language is not English. Keeping up with technological, economic, social, cultural and political advancements in the first world is becoming an extremely laborious and costly, and ultimately impossible task. In the Arabic world, there has been quite a great deal of awareness of the need of Arabization in all walks of life. Nevertheless, the task today has become overwhelming, human translation seriously fails to cope with the pace and scale of information production, especially in English. This is why the need and importance of MT should be stressed. This area has been the target of governmental, academic and private institutions in the West and has undergone a lot of development, helped by an important effort to advance the state of CL. In the Arab world, unfortunately, research and development of MT and CL for Arabic has remained limited with almost no involvement of governmental institutes to support it. This is probably due to a lack understanding of the full potential of MT, NLP and AI more generally (Zantout and Guessoum, 2000: 118).

Arabs should realize the importance of MT in relation to their language and their culture. MT from other languages into Arabic is obviously crucial. However, many researchers indicate that MT into Arabic still faces many serious difficulties some of which are listed below (AL-Sieny, 2000:174):

1- Multi meanings of a single lexical item in the SL.
2- Homophony.
3- Terminology and idiomatic expressions.
4- Metaphor.
5- Syntactic problems (e.g., subject/verb agreement; attribute/noun; number; gender, etc).
6- Problems relevant to the complexity of construction in the ST.
7- Aesthetic standards of the ST.
8- Interference of foreign terminologies.
9- Variation of the cultural environment between the SL and the TL.

MT of Arabic into other languages is also an important area since it enables Arabs to disseminate information more efficiently and make their culture and products accessible to more people. It is worth mentioning here that we do have products to translate from English-into-Arabic, but we still do not have reliable products to perform the opposite, i.e., Arabic-into-English. The reason behind this is that the linguistic analyzer for English is already built. The analyzer for Arabic is still subject to extensive research and the most common tool utilized to achieve this is the Lexical Functional Grammar, to build the Arabic language model. Some problems faced here are word ambiguity, text semantics, the absence of one-to-one lexical relation and word-to-phrase translation (ACP Review, 1996: 1-2).

2.12.1 Existing Arabic MT Systems and Arabic MTE

Compared to its American, European and Asian counterparts, Arabic MT is still in its infancy stage. Nevertheless, the area has recently been gaining momentum. As first fruits of this endeavor, commercial software has been produced by companies such as Apptek, ArabTrans, CIMOS and ATA (See 2.11.4).

However, despite the existing efforts in the area of Arabic MT, Arab efforts towards developing the research in such a vital area have been minimal (Zantout and Guessoum, 2000; 135).
As to the evaluation of Arabic MT systems, work remains non-existent (Jihad, 1996:38-48). Only very brief surveys of a number of Arabic MT systems including: Transfer, ArabTrans and AL-Wafi 2.00 have been presented (ibid. and Qendelft, 1997). Nevertheless, these brief assessments were carried out in a very ad hoc way and thus failed to convey any scientifically convincing conclusion as to the quality of the evaluated system. (Recently, a number of Arabic MT evaluation research have been carried out by Zantout and Guessoum to be presented in section (2.14) of this chapter). In fact, evaluation of Arabic MT is still very sparse and very much non-systematic. The few attempts at evaluating Arabic MT systems, did not rely on any firm or formal evaluation methodology. Despite such ad hoc methodology, the evaluations have shown some serious shortcomings of the available Arabic MT systems. For instance, a very brief evaluation of AL-Wafi (Anon,1996:27-28) was presented. Examples of English-into-Arabic translations performed by AL-Wafi were given. The results were quite disappointing (See also AL-Wasiti, 2000: 65-79). There are still serious contextual reasoning problems with getting the word order or verb tenses right, choosing the right word or expression out of a number of alternatives or translating ambiguous sentences. As a result of a number of drawbacks, translation done by AL-Wafi was shown to produce texts which are not faithful or accurate with respect to the ST, and which are often quite unintelligible. Actually, the output reflected many deficiencies in translating even basic computer terms. Three Arabic MT systems (AL-Wafi 2.00, AL-Mutarjim AL-Arabey 2.00 and ArabTrans) were recently evaluated. The evaluation was based on a methodology developed in Guessoum and Zantout (in preparation). The results as detailed in AL-Sikhan, Zantout and
Guessoum (1999) show very severe shortcomings especially in relation to the grammar and meanings of the translated sentences.

However, Zantout and Guessoum (2000:137) confirm that although the evaluation of these systems may sound very pessimistic, it is meant to give all researchers and developers interested in Arabic MT the right dose of realism so that a change is initiated. It is true that the 15 years or so in the life of Arabic MT is a minor in the life of a discipline. Nevertheless, it is important to realize now, and before it becomes too late, that major efforts are urgently needed (this is also indicated in Guessoum, 2002a).

2.13 Arabic MT and the Internet

The Internet and the World Wide Web have become an integral part of everyday life, an important source of information and a communication medium. For on-line translation, there are basically two categories: a.) the provision of services for the translation of technical, scientific, operational documents, etc. for large companies and b.) the on-line production of translations of short, informal texts.

Schenker (2001:1-2) maintains that the quality of the Internet free translation services tends to be uneven because the subject matter is so diverse and the process relatively imprecise. It is called “gisting” (8), because it yields translations that provide the gist of a document (See also Transparent Language, 2002).

Zantout and Guessoum (2001:336) state it is unreasonable to ask non-Arabs to publish information in Arabic in order to make the Internet accessible to Arabs. It is the responsibility of Arabic speaking users to develop tools that will make the Internet more Arabic-friendly. Using human
translators to perform such a task is not a practical choice. This is due to the amount of information available on the Internet and the fast rate at which it is growing. Human translation is not appropriate since it will take hours for a dedicated human translator to be able to produce a translated web page that corresponds exactly to the source the user requested.

Certain on-line web sites to translate web pages and different text types and documents from English-into-Arabic are available today. Zantout and Guessoum (ibid:340) indicate that translating web pages automatically is not a trivial task and there are several possible approaches for web page translation. Pure texts can be translated using commercially available MT systems on the web. Several basic translation systems, as well as electronic dictionaries, are available free of charge on-line.

The following are commercially-available systems for on-line translation between English and Arabic or the reverse (Computers and Arabic, 2002):

1. **AL-Misbar (ATA Software Technology)**: This web site translates on-line texts freely from English-into-Arabic. Arabic translation software is powered by AL-Mutarjim AL-Arabe and AL-Wafi MT engine (AL-Misbar Net, 2000). The translation is performed by entering a web site address to translate Uniform Resources Location (URL) or by entering an English text. The translation options offered by the system are: a.) translation theme, b.) transliteration of abbreviations, c.) transliteration of proper nouns and d.) showing vowels (i.e., diacritics). It is accessible at www.ALMisbar.com.

2. **Cimos** using Multi–Lingual Translation Service (MLTS) Internet translation service. MLTS is a software package that rapidly produces high quality translation. Rendering the target text understandable with high
accuracy, MLTS integrates: 4 dictionaries (general, idiom, specific dictionary and user dictionary), MT database, verb conjugation package, lexical analyzer, syntactic analyzer and semantic analyzer. A wide range of subject areas such as science, commerce, computer, medicine, banking, etc. are covered. It has exactly the same characteristics and requirements of An-Nakel AL-Arabi that goes from English-into-Arabic and in reverse (Translation Engines, 2002). It is accessible at www.cimos.com.

3. Ajaeb: According to Bell (2002), Sakhr has released a free application which provides a translator and dictionary via the web. It offers Arabic-English-Arabic word translation, English-Arabic sentence translation, and English/Arabic translation to French, German and Turkish.

Ajaeb’s most useful feature is its ability to translate entire sentences from English- into- Arabic. First, the user has to select (highlight) the sentence in Word or any other Windows programme, hold down the <Ctrl> key and right click on the sentence.

This opens Ajaeb, which spell-checks the text and provides an Arabic translation. A sentence can be typed directly into the translator too. When the user looks up an English or Arabic word, Ajaeb displays its dictionary definition and synonyms in the specified TL.

Ajaeb also makes it easy to search the Internet for the word or phrase typed by the user. The application processes search queries using Sakhr’s web dictionary and the IDRISI bilingual search engine.

Ajaeb makes errors of different types and it mistranslates many words and expressions. The programme is very user-friendly and the speed is acceptable with a 56 Kbs Internet connection, it took 20 to 30 seconds to translate each English sentence.
The main weakness of Ajaeb is that the user must be connected to the Internet when using it. The application’s translation engine is not hosted on the user’s PC, and all of the work is done by Sakhr’s servers.

Ajaeb is available for download from Sakhr’s Arabic/English portal, accessible at www.ajeeb.com or http://eajaeb.ajeeb.com. It is a convenient application suitable for small translation tasks.

It should be mentioned here, that due to restrictions imposed on access to this site to Internet users and to SAKHR Web Site in general; the researcher of the present study has not been able to enter this Web Site or to test its translation engine.

4. Tarjim : It is an English/Arabic MT engine developed by Sakhr Software. Sakhr Software has been working in Arabic NLP for about 12 years, and in English NLP for about 4 years now. Three years ago, they initiated their Bidi English/Arabic MT project (Transfer Model). They now have version 1 of the MT engine that has been used to develop Tarjim.com.site, which is the first Web on-line translation (English-into-Arabic) on the Internet. The average accuracy is 60% (Translation Engines,2002). The system is accessible at www.Sakhr.com.

As any web site using Sakhr’s own MT systems, this web page has been inaccessible to the researcher of the present work.

Now, since access to certain English-into-Arabic translation web sites (e.g., Ajaeb and Tarjim) has been denied, and as other web sites (e.g., AL-Misbar and MLTS) use the same commercial MT systems available in the market for on-line translation (e.g., AL-Wafi; AL-Mutarjim AL-Arabey and An-Nakel AL-Arabi), the researcher of the present study finds the evaluation of on-line MT systems outside the scope and objectives of this work. In
addition, the translation of web pages and on-line translation has its own methods and approaches (See Zantout and Guessoum, 2001: 333-57), so the evaluation which is restricted to this type of translation is left for further research.

Above all, for limits of time and space, and since in most cases, on-line translators give the gist of translation, the researcher has excluded this aspect of the present work.

2.14 Recent Research on MTE

Despite the fact that non-Arabic MT systems still display weaknesses and the TTs they output still require fairly substantial post-processing, their state is by far better than that of Arabic MT systems. Moreover, despite the existing efforts in the area of Arabic MT, Arab efforts toward developing the research in such a vital area have been minimal. As for the evaluation of Arabic MT systems, attempts are still very shy and very much non-systematic.

In contrast to the Arabic MT and MTE research projects, there have been considerable continuity of research and development projects and workshops in America, Europe and Japan (some of the well-known early attempts have been mentioned in 2.5.1 of this study). The ISLE Project (1999-2002) has organized a series of workshops and conferences on MTE which concentrated on practical work using the taxonomy. Because of the limitation of space, the researcher intends here to shed light on the most recent activities in the series offering a number of interesting examples of using the taxonomy in practice and other relevant issues.

Only some examples picked out of the papers presented at the MT Summit VIII in Santiago/Spain (18-22 September, 2001) are given in this
section, since they cover a very wide range of topics closely related to the present study, including: a.) the development of new metrics, investigations into possible correlations between metrics, ways to take into account different user needs, novel scenarios both for the evaluation and for the ultimate use of the MT system and ways to automate MT evaluation.

2.14.1 Arabic MTE Research

2.14.1.1 MTE Study in Iraq

“A comparative study of Machine Translation Systems” is the title of an unpublished M.A thesis in computer science carried out by AL-Bahadily (2001). It is the first study in Iraq regarding the evaluation of MT systems that mainly aims at comparing some available MT systems (AL-Wafi 2.00, AL-Mutarjim AL-Arabey 2.00 and An-Nakel AL-Arabi 2.00) with respect to errors and problems in translation. The findings of the study show the merits and demerits of each MT system and the researcher has designed a checklist to be used in the evaluation of any MT system which can be customized and improved according to the evaluation purpose.

Some linguistic and computational criteria were considered in the evaluation, such as: translation of terminologies, semantic analysis, translation of abbreviations, dictionary, multi-meanings, pronouns, speed, protection of system, documentation of system, size of TTs, type of files, etc.

It is noticeable that despite the intensive efforts of the researcher, the evaluation does not follow the international standards, taxonomies, metrics and methods.

2.14.1.2 Very Recent Arabic MTE Studies
Guessoum and Zantout (2000:135 and 2002a.) have confirmed that, “to the best of our knowledge, work [on Arabic MTE] remains non-existent.” Since they have some recent research based on the international standards, taxonomies, metrics and methods, they can be considered Arab pioneers in the field of MTE.

In addition to other research papers on MT and on-line translation, Guessoum and Zantout have two researches on MTE in 2001 and 2002. They believe that subjectivity of NLP/MT system evaluation stems from the fact that evaluation methods rely heavily on humans, i.e., the assessment will depend on the evaluator’s background and skills. Hence, more objective evaluation methodologies need to be developed. So, in their studies, they tried to automate some evaluation tasks in order to minimize the amount of subjectivity in the evaluation process.

In the first paper, Guessoum and Zantout (2001) presented a methodology for automating the evaluation of the grammatical coverage of MT systems that go from English-into-Arabic (e.g., AL-Wafi 2.00; AL-Mutarjim AL-Arabey 2.00 and ArabTrans). The methodology was based on the importance of unfolded grammatical structures, which presented the most basic syntactic pattern for a sentence in a given language. A database of unfolded grammatical structures was built to evaluate the parser of any NLP or MT system. The evaluation resulted in an overall measure called the grammatical coverage. The evaluation methodology was implemented and tested with the three Arabic MT systems. The evaluation results showed a poor grammatical coverage for all three systems, confirming the disappointing output frequently produced by such systems. The evaluation also indicated that two systems (i.e., AL-Wafi 2.00 and AL-Mutarjim AL-Arabey 2.00), produced by the same company, used the same parsing
engine. The methodology could also be extended to the evaluation of the semantic correctness, pronoun resolution correctness and style of TT.

In their second paper, the researchers introduced a new methodology devised to enable developers and users of MT systems to evaluate their lexicons semi-automatically. This new methodology was based on the idea of the importance of a specific word, more precisely, word sense, to a given application domain. The importance, or weight, determined how the presence of such a word in or its absence from the lexicon affected the MT system’s lexical quality, which in turn would naturally affect the overall output quality. The method, which adopted a black-box approach to evaluation, was implemented and applied to evaluating the lexicons of three commercial English–Arabic MT systems (AL-Wafi 2.00, AL-Mutarjim AL-Arabey 2.00 and ArabTrans). A specific domain was chosen in which the various word-sense weights were determined by feeding sample texts from the domain into a system developed specifically for that purpose. Once this database of word sense and weights was built, test suites were presented to each of the MT systems under evaluation and their output was rated by a human operator as either correct or incorrect. Based on this rating, an overall automated evaluation of the lexicons of the systems was deduced (Guessoum and Zantout, 2002:127-49).

In the University of San Francisco, EL-Nady (2001) investigated native speaker’s assessment of the adequacy and acceptability of MT. An-Nakel AL-Arabi software was used by the researcher to translate selected Arabic and English opinion articles from ‘AL-Magala’ and ‘Time’. An assessment sheet was designed by the researcher and was used by ten native speaker linguists in their evaluations of the MTs.
The native speaker’s assessment of the MTs showed that the Arabic and English translations could not be adequate nor acceptable to the native speakers of the language of translation, not even to the translators as a first draft. The challenge of the translation, however, did not originate from the use of the MT software, but from the articles themselves due to language and cultural barriers.

An-Nakel is found to be a useful training or educational tool for those who work in the field of translation and interpretation. With software upgrading, especially the ability of this software to learn new vocabularies and rules, and the possibility of using it as a tool by skilful translators, it appears that the potential of MT in multi-cultural media is good, particularly if the writers, when they write their articles, take into their consideration the readers of other cultures.

2.14.2 Review of Non-Arabic MTE Research

Research and development in the area of MT for non-Arabic systems has been an active discipline for almost half a century. In 1949, Warren Weaver suggested that computers could be used to perform text translation. Since then, non-Arabic MT has gone through a lot of development and has benefited of all the research and products of NLP. MTE has also been given considerable attention especially in recent years. In fact, this is the prime reason for the existence of a wide gap between non-Arabic MT and Arabic MT. This is because MT(E) has been selected as a ‘strategic choice’ by the officials in the USA, the European Union, Japan and Canada. Such an awareness of the crucial and strategic nature of MT(E), is still lacking in the Arab world. In fact, the existing Arabic work in this field is mainly based on individual researchers in universities and research
centers or on commercial companies, many of which happen to be located in the West.

The evaluation of MT and NLP tools has been accepted as a sub-area with its own right in America, Europe and Japan. The examples below reflect the advanced and systematic level of MTE research they have reached. The research papers are presented on the basis of three sub-section headings: a.) Evaluations of MT aspects and MT output quality, b.) ISLE-based evaluations and c.) Automatic evaluations of MT systems. They are listed according to the alphabetical order of their authors.

2.14.2.1 Evaluations of Different MT Aspects and MT Output Quality
In a research paper presented in 2001, Darwin described a small-scale but organized attempt to evaluate output quality of several Japanese MT systems. The project also served as the first experiment of the implementation of the in-house MT evaluation guidelines created in 2000. Since time was limited and the budget was not infinite, it was launched with the following compact components: five people, 300 source sentences per language pair, and 160 hours per evaluator. The quantitative results showed noteworthy phenomena. Although the test materials were presented in a way that evaluators could not identify the performance of any particular system, the results were quite consistent. The scoring ratio that the two E-to-J evaluators employed was almost identical, while that of J-to-E evaluators was similar. This indicated that high-quality output had universal appeal. Additionally, the evaluators noted that stronger systems, regardless of language pair, tended to be superior in source sentence analysis, target sentence arrangement, word choice and lexicon entries, whereas weaker systems tended to be inferior in these areas. As for language-pair
comparison, the results indicated that English-to Japanese systems require more improvement than their counterparts, judging from the scores given and the number of unfound words recorded.

EL-Hadi (2001), in France, examined some of the problems encountered in designing an evaluation for an MT system. The ST, in French, dealt with biotechnology and animal reproduction. It was translated into English. The output of the system (i.e., the result of assembling of several components), as opposed to its individual modules or specific components (i.e., analysis, generation, grammar, core, lexicon, etc) was evaluated. The evaluation concentrated on translation quality and its fidelity to the ST. It was not a comparative evaluation. It was only a test of a specific MT system.

In Korea, Koh et al. (2001) described KORTERM’s test suite and their practicability. The tests-sets were constructed on the basis of fine-grained classification of linguistic phenomena to evaluate the technical status of English-to-Korean MT systematically. They consisted of about 5000 test-sets and were growing. Each test-set contained an English sentence, a model Korean translation, a linguistic phenomenon category, and a Yes/No question about the linguistic phenomenon. Two commercial systems were evaluated with a Yes/No test of prepared questions. Total accuracy rates of the two systems were different (50% vs. 66%). In addition, a comprehensive test was carried out. The researchers found that one system was more comprehensible than the other system. The results seemed to show that their test suite was practicable.

Maier, Clarke and Stadler (2001) described the evaluation of MT system for use in a large company. To take into account the specific requirements of such an environment, a pragmatic approach for the evaluation was developed. It consisted of five steps ranging from a specification of the
evaluation process to the integration of the chosen MT system in a given infrastructure. The process included a specification of MT evaluation criteria relevant to systems which had to be employed for a large customer base. The paper also showed the results of such an evaluation study which was recently carried out at Corporate Language Services (CLS) AG, where COMPRENODIUM is in the meantime being employed as corporate MT system.

Reeder et al. (2001) reported the results of an experiment in MTE, designed to determine whether easily/rapidly collected metrics could predict the human generated quality parameters of MT output. In this experiment, they evaluated a system’s ability to translate named entities, and compared this measure with previous evaluation scores of fidelity and intelligibility. There were two significant benefits potentially associated with a correlation between traditional MT measures and named entity scores: a.) the ability to automate named entity scoring and thus MT scoring and b.) insights into the linguistic aspects to task-based uses of MT as captured in previous studies.

In the University of Zurich, Volk (2001) found new methods of comparative evaluation to compare between several commercial MT systems for the same language pair. He proposed a black-box method for comparing the lexical coverage of MT systems. The method was based on lists of words from different frequency classes. It was shown how these word lists could be compiled and used for testing. He presented the results of using this method on six MT systems that translate between English and German.

2.14.2.2 ISLE-Based Evaluations

Evaluation guidelines for a given domain or task must be rooted in a general model for software evaluation.
Following the guidelines for MT evaluation proposed in the ISLE taxonomy, Bruckner and Plitt (2001) presented considerations and procedures for evaluating the integration of machine-translated segments into a larger translation workflow with translation memory (TM) systems. The scenario focused on the software localization industry, which already used TM systems and looked to further streamline the overall translation process by integrating (MT). The main agents involved in this evaluation scenario are localization managers and translators; the primary aspects of evaluation were speed, quality and user acceptance. Using the penalty feature of TM systems, the authors also outlined a possible method for finding the “right place” for MT produced segments among TM matched with different degrees of fuzziness.

Marrafa and Ribeiro (2001) reported the first results of an on-going research on evaluation of MT quality. The starting point for this work was the framework of ISLE, which provided a classification for MTE. In order to make a quantitative evaluation of translation quality, the researchers pursued a more consistent, fine-grained and comprehensive classification of possible translators and they proposed metrics for sentence level errors, specifically lexical and syntactic errors.

Miller and Vanni (2001) focused in their study on the selection of MT output features suggested in the ISLE framework to develop metrics for the multi-dimensional characterization of MT quality. They defined each metric and described the rationale for its development. They also discussed several of the finer points of the evaluation measures that arose as a result of verification of the measures against sample output texts from three MT systems.
Popescu-Belis, Manzi and King (2001), in their study, considered as a starting point the ISO/EAGLES guidelines for NLP software evaluation. From these considerations, they derived several principles for a taxonomy aimed at the evaluation of machine translation systems. Then, they compared two editions of such a taxonomy, arguing in particular for a dichotomy relating user needs to system characteristics. They also outlined the software infrastructure underlying the electronic publication and updating of the taxonomy, and concluded with a brief overview of the workshops that were staged to test, modify and disseminate the taxonomy.

2.14.2.3 Automatic Evaluations of MT Systems
Akiba, Imamura and Sumita (2001) addressed the challenging problem of automatically evaluating output from MT systems in order to support the developers of these systems. It proposed an automatic ranking method that, by using multiple edit distances, encoded machine-translated sentences with a rank assigned by humans into multi-dimensional vectors from which a classifier of ranks was learned in the form of a decision tree (DT). The proposed method assigned a rank to MT output through the learned DT. The proposed method was evaluated using transcribed texts of real conversations in the travel arrangement domain. Experimental results showed that the proposed method was more accurate than the single-edit-distance-based ranking methods, in both closed and open tests. Moreover, the proposed method could estimate MT quality within 3% error in some cases.

The main goal of the work presented by Rajman and Hartley (2001) was to find an expensive and automatable way of predicting rankings of MT systems compatible with human evaluations of these systems expressed in the form of fluency, adequacy or informativeness scores. The researchers’
approach was to establish whether there was a correlation between rankings derived from such scores and the ones that could be built on the basis of automatically computable attributes of syntactic or semantic nature.

Also in 2001, Reeder aimed at testing the efficiency of applying automated evaluation techniques, originally designed for human second language learners to MT system evaluation. The researcher believed that such evaluation techniques would provide insight into MTE, MT development, the human translation process and the human language learning process. The experiment described only intelligibility of MT output. The evaluation technique was derived from a second language acquisition experiment that showed that assessors could differentiate native from non-native language essays in less than 100 words. The researcher duplicated this experiment to see if similar criteria could be elicited from duplicating the test using both human and MT outputs on the decision set.

Approaches to the automation of MTE were attempted by White and Forner (2001) to connect some rapidly measurable phenomenon with general attributes of the MT output and/or system. In particular, measurements of the fluency of output were often asserted to be predictive of the usefulness of MT output in information-intensive, downstream tasks. The connections between the fluency (intelligibility) of translation and its informational adequacy (fidelity) were not actually straightforward. This paper discussed a small experiment in isolating a particular contrastive linguistic phenomena common to both French-English and Spanish-English pairs and attempted to associate that behaviour in machine and human translations with known fidelity properties of those translations. The results showed a definite correlative trend.
In Japan, Yokoyama et al. (2001) proposed a quantitative evaluation of MT system. The method was roughly that an example sentence in Japanese was machine translated into English, and into Japanese using several systems, and that the comparison of output Japanese sentences with the original Japanese sentence was done for the word identification, the correctness of the modification, the syntactic dependency and the parataxis. By calculating the score, the researchers could quantitatively evaluate the English MT. However, the extraction of word identification, etc. was done by human, and the fact affects the correctness of evaluation. In order to solve this problem, they developed an automatic evaluation system.

It is worth mentioning that a workshop at the LREC (May, 2002) conference concerned itself with the automatic evaluation of MT systems, “Human Evaluators Meet Automatic Metrics”. The full workshop data can be downloaded from: http://www.issco.ch/projects/isle/mteval-may02/. Also, a recent workshop on the evaluation of MT, “Towards Systematizing MTE “, was held in New Orleans, Louisiana, USA (September, 2003). Researches can be downloaded from: http://www.amtaweb.org/summit/W52/MT_Summit_IX_Workshop.htm.

2.15 Summary
Over the last ten years, the level of interest and concern expressed by the NLP community with regard to evaluation has increased substantially. Now, it has become widely accepted that work in NLP should pay close attention to the evaluation of certain characteristics (e.g., coverage, speed, correctness) of some NLP/MT systems or one of its parts. Two or even more
MT systems might be compared to indicate which one provides the better results, or some metrics might be developed by means of which the performance and improvement of a given system can be determined.

This chapter has focused on why evaluation takes such a pre-eminent role in the field of MT. It has been a highly visible aspect of the promise andfailings of MT since its beginnings. It is profoundly important, because of the investment that goes both into development of MT approaches, and adoption of MT processes in a work environment. Besides, evaluation is very difficult, because translation does not generally permit comparison against single standards, and because the variety of uses and users require particular investigations of particular properties of MT systems.

In addition to these and other issues relevant to MTE, the chapter has discussed the different types of evaluation which are intended to tell us what we need to know about MT systems at different points in the system’s life, and from the perspectives of the people who must use it or make decisions about it. However, the new trend in MTE point to the need for the development of evaluation methodologies that minimize the amount of subjectivity as much as possible. It is believed that the best way to achieve this goal is to automate, as much as possible, any evaluation task or part thereof.

In fact, whatever new method may emerge, and whatever methods will ultimately be unnecessary, it is clear that evaluation will remain very near the center of MT awareness.

More attention to the model of evaluation (within which the aspects of the system to be measured and the appropriate measures to be used are classified), the focal part of this work, will be given in the next chapter. It should be noticed that in addition to some computational criteria (e.g.
storage, flexibility and portability) concentration will be on the system’s external characteristics in the ISLE taxonomy 3. As for the type of evaluation, the black-box approach, adequacy/declarative and comparative evaluations will mostly be applied.

Notes to Chapter Two

(1) A radar plot resembles a spider web, in which each radial axis represents one feature and the single circular correcting thread intersects each feature (axis) at the point corresponding to its value. A good system would score highly for all features, thus having the connecting thread circle far from the center; a bad system’s connecting thread would circle the center closely.

(2) According to White in Somers (in press), interactive testing refers to tests designed to make sure an improvement in one area actually works and does not adversely affect another area).

(3) In broad terms, we may distinguish between systems which attempt to translate whole texts without intervention by human operators and these which require assistance to resolve problems of ambiguity in the ST or to select the most appropriate word or phrase in the TL. The former type is referred to as “batch” systems because the whole text is processed as “one-task produced “ output requiring revision by a human translator to a greater or lesser extent. This revision (or post-editing) may well be substantial if the text is intended for publication, but it may be minimal if the text is intended only for information scanning purposes within an organization. An alternative mode of operation with such systems is to ‘pre-edit’ the input texts to reduce ambiguities and complexities of structure. The other type of MT
systems is “interactive” systems where the computer seeks the assistance of a human operator during the translation process itself in order to resolve ambiguities and make decisions which may be difficult for programmers to define rigorously (Hutchins, 1991:1-3).

4) According to Arnold (1994:4), a controlled language is referred to as a specially simplified version of a language which is adopted (typically by a company or a documentation section of a company) as a partial solution to a perceived communication problem. Both the vocabulary and the syntactic structures may be restricted.

5) TAUM METEO is an MT system used by the Canadian Meteorological Office for translation of weather bulletins from English to French. Grammars for each stage of the translation process are expressed in the same formalism. METEO constitutes a paradigm case of simplicity in design, which for the limited domain of texts it handles, results in particular effectiveness (Kilby, 1995:3).

6) The discussion of different types of MT systems will be incomplete if the concepts of translation for assimilation and translation for dissemination are not mentioned (Guessoum and Zantout, 2000:125). In the former, translation is targeted towards giving the user a summary of the most important ideas in the TT. The later, would be more concerned with producing a TT that would have a good style and would be grammatically and syntactically correct in the TL. It is, therefore, beneficial to consider the purpose of translation when evaluating the output. The first purpose can tolerate low quality translations, while the second requires the system to deal with the highest available quality.
(7) As Sager (1994:269) indicates, MT systems following the transfer approach are more complex than others which follow the direct or interlingual approaches in translation. In the analysis phase, the SL text is processed to the depth required by the rules of its grammar and the result is expressed in a notation suitable for the transfer phase. In the transfer phase, this representation, based on the SL analysis, is transformed into a representation suitable for the generation of a TL text. Whereas the analysis phase does not take into account any feature of the TL, the transfer phase can make adjustments to the results of the analysis with regard to the generation requirements for the TL. In the generation phase the transfer representation is then transformed into a text in the TL without back-reference to the results of analysis (For further details, see Shuttleworth, 1997:176-77).

(8) The emerging global connectivity via the Internet, Intranets and other networks has created an urgent need for inter-language communication tools that provide immediate “gisted” translations. “Gisted” translations are draft-quality translations that provide individuals and professionals with the meaning of the original foreign-language document so that they can determine the relevance to themselves or their business. “Gisted” translation may be the only solution to overcome a language barrier. Despite limitations, MT systems will become an important and indispensable tool for businesses and individuals where global communication is a priority (Transparent language, 2002).
Like other tools, MT systems must be evaluated. Their effectiveness, adequacy and cost need to be assessed in order to discover how well they are meeting their user’s requirements. Methods of evaluation must themselves meet criteria of efficiency and cost-effectiveness; they should provide useful information at reasonable expense.
With regard to the efficiency of MT, it is useful to distinguish between macro-(or global) and micro-(or detailed) evaluation. In the latter, evaluators concern themselves with diagnosing the causes of errors, proposing improvements and measuring their effectiveness. As such, it requires lengthy and painstaking work and is therefore outside the scope of the present study. Instead, this project concentrates on macro-evaluation of a system’s performance.

Since tests of the quality of MT have often been based upon extremely subjective factors, it is essential to decide upon several detailed criteria for the evaluation and then for the evaluation to be performed strictly on the basis of those criteria.

Current general purpose commercial MT systems cannot translate all texts reliably. Output can sometimes be of very poor quality indeed. Because of its enormous influence on the overall translation cost and effectiveness, translation quality is a major aspect in MTE.

In this chapter, the researcher will look at the kind of thing that should matter to potential users and evaluators of MT systems. Then, the model of evaluating the MT systems in the present study will be discussed. This involves identifying and explaining the types of methods, criteria and metrics relevant to the aims of the current evaluation. In addition, discussion on evaluation requirements, translation quality and other evaluation aspects will be presented.

Finally, the ISLE taxonomy 3 for the evaluation of MT (FEMTI, 2003) will be used as the basis for the evaluation task in this study. This evaluation concentrates on the second part of the ISLE’s taxonomy regarding the quality model, and its six system external quality characteristics. The object of the evaluation is to compare three English/Arabic MT systems and to judge the quality of their output in the light of the above-mentioned taxonomy and the EAGLES (1999) 7 major steps of MTE (See 2.5.1.3 in this study).

3.1 Evaluation of MT Systems: Background

Software evaluation is proving ever more crucial with the frenzied growth of software development. No matter what the domain is, software evaluation can target the quality of the software under assessment, its cost, and/or performance. Guessoum and Zantout (2002:127) indicate that “testing” and
“maintenance” are two major parts of a software life cycle (See Fig. 3.1 below). These are directly relevant to software evaluation, performed by the software developers during the development process. The additional aspects of software evaluation are performed after the software has been developed and may be carried out by potential users and/or researchers. Moreover, any software can be evaluated for its intrinsic features, or for purposes of comparing them to those of similar software so as to select the most suitable amongst them in terms of the features that are targeted by a given institution or individual.

Figure (3.1). V-diagram Showing the Place of Evaluation in Software Development.

A serious issue to consider when dealing with an MT system is the ability to evaluate it “scientifically”. This is in order to establish its usefulness and compare it with other already available systems. This evaluation should cover both its computational and linguistic features and should be informative, especially to potential users of the system, as to its capabilities and updatability (ibid : 130).

In terms of quality assessment, the evaluation of NLP systems has been divided into two main approaches: “glass-box” and “black-box” evaluations (Hutchins and Somers, 1992 and
Nyberg et al., 1994). In the latter, the evaluator has access only to the input and output of the system under evaluation (an approach followed in the present study). In the former, the evaluator also has access to the various workings of the system and can thus assess each subpart of the system independently and in association with the others. In addition, NLP systems need to be evaluated for the limitations of the system, its benefits, performance, stability and portability among other criteria.

Since MT systems combine lexical analyzers, morphological analyzers, parsers, semantic disambiguation modules, generators, pragmatic modules, etc., it is important to be able to evaluate these various components individually as well as to evaluate the overall system.

In general, evaluation is clearly a complex task since it needs to take care of all the aspects mentioned above. As Guessoum and Zantout (2002:128) maintain “it is even more complex in the case of MT systems.”

Indeed, the evaluation of MT systems introduces a number of additional complications, chief of which is its currently subjective nature. The authors believe that such subjectivity comes from the fact that all current evaluation methods rely heavily on humans to produce grades that are used to evaluate the system under consideration. The assessment of the quality of some system (or component) output may depend on the evaluator’s background, skills or even taste. In addition, the evaluation process will be affected by the proficiency (or lack of proficiency) of the human evaluator in the various facets of the languages which are involved in the translation. Besides, the difference between a score of 3 or 4 on a scale of 5 for some feature of the system under evaluation may not be that obvious for various evaluators. Even when strict and very clear rules are introduced at the beginning of the evaluation process, the exact number given to a system component under evaluation would vary from one human evaluator to another. Also, various MT systems may use different approaches to translation and may be used in application domains or settings that may make their performance judged differently. Further, given that the evaluators do not have access to the internal workings of the system under evaluation, and are therefore forced into “black-box” evaluation, their interpretations of their evaluations do not necessarily yield appropriate diagnoses of the observed errors (ibid.).

In the previous arguments, the researcher explains why the evaluation of an MT system’s quality is far from being a simple task. Moreover, one should note that MT systems most of the time give ‘raw’ output whose quality necessitates post-processing by human translators. Thus, the evaluation of the performance of an MT system will also depend on the acquaintance of the human translators with the system. Besides, the overall cost of using a given MT system should
also take into account the cost introduced by the use of a human post-processor (and sometimes even pre-processor).

So, there is a need for the development of evaluation methodologies that minimize the amount of subjectivity as much as possible. Recently, it is believed that the best way to achieve this goal is to automate, as much as possible, any evaluation task or part thereof.

### 3.2 Levels of Evaluation

MTE is carried out at two levels and for two main purposes: a.) macroevaluation (global/total evaluation) and b.) microevaluation (detailed evaluation). These are discussed in detail in the following sub-sections (See Fig. 3.2).

#### 3.2.1 Macroevaluation

This level of evaluation concerns itself with the assessment of the system’s overall performance (Van Slype, 1979a: 88). It deals with all the criteria and all the methods used or proposed to assess the “static” quality of an MT system, i.e., its quality at the moment of evaluation, and regardless of the manner by which this quality has been reached. It aims at examining the acceptance of a translation system, comparing the quality of two translation systems or two versions of the same system and/or assessing the usability of a translation system (Van Slype, 1979b: 12 and 21).

According to Bennett et al. (1986: 95) and Sager (1994: 146), if evaluation occurs at a general macro level, it is broadly qualitative and can be directed at a comparative appreciation of products with respect to cost, time and adequacy for the intended purpose, which comprises such criteria as acceptability, intelligibility and even reading time by the end user.

Macroevaluations serve the purpose of assessing the relative merits of one system compared with another. They have to judge output of competing systems with respect to the usefulness for the intended purpose for each type of application, be this post-editing, direct use, pre-translation etc. Macroevaluations are also the first step for comparing time and cost factors between human and MT (Sager, ibid: 264).
Van Slype (1979b:56) points out that the macroevaluation of a system is the operation which consists in assessing the manner in which the system answers the requirements and the needs of its users, actual or potential, regardless of what occurs inside the “black-box”. It has the purpose of measuring the adequacy of the output from the system to its environment, without seeking to diagnose the causes of its inadequacy, if any, and without pinpointing the component(s) that could usefully be modified to improve adequacy. He adds that the field of macroevaluation is limited:

a. on the one hand, by the marketing, as regards the extent to which a product or a service meets the market demand.
b. on the other hand, by microevaluation, which is concerned with the diagnosis of errors and with improvability.

Further, it would be possible to envisage establishing a classification of the techniques of macroevaluation on two separate levels:

a. a list of criteria (e.g., intelligibility, fidelity, etc.).
b. a list of methods of measuring these criteria (e.g., cloze test, rating of sentences on 9-point scale, etc.).

In fact, criteria methods can be used to measure the value of several distinct criteria. Also, the measuring methods are associated with the criteria, i.e., for each criterion, the method(s) practiced and proposed by the various authors are presented. In certain cases, the same method can be used to measure different criteria (ibid.).

### 3.2.2 Microevaluation

Microevaluation is used or proposed to assess the “dynamic” aspect of an MT system, i.e., its scope for improvement and the limits thereto. In other words, it concerns itself with the analysis of the errors attributable to the system and the means of avoiding them with a view to assessing the extent to which the system can be improved and establishing an improvement

While macroevaluation is usually directed to alternative products by different people or produced by different methods, microevaluation can concern itself with comparison among alternatives and evaluation of a single product. At the micro level, evaluation is concerned with criteria of correctness in terms of choice, order and grammatical agreement of items. Microevaluation is essential for the improvement of translation performance and MT systems (Sager, 1994: 146).

Van Slype (1979b:116) maintains that although there are far fewer microevaluation studies in the field of MT, both in the literature and in practice than there are macroevaluation analyses, some authors stress the importance of them. Sinaiko (1978:7), for example, recommends that the evaluations should aim at discovering why the MT system examined is inadequate. Similarly, Vauquois (1978:2) recommends analyzing the nature of the faults submitted to the reviser, discovering their cause and conducting an assessment of the extent to which simple modifications to the system would enable them to be avoided. These authors’ contributions to the microevaluation discussion can be classified into five groups according to the level of the analysis, i.e., grammatical level, formal level, caused level, remedy level and improvement level. The last three are in effect the only ones, which correspond to the definition of microevaluation. In this respect, Sager (1994:265) indicates that microevaluation can have several functions. It can be ‘diagnostic’, i.e., examining the cause of output error or ‘prognostic’, i.e., assessing errors for their potential solutions and future avoidance.
In short, having presented a detailed discussion of each level of evaluation, it seems obvious that macroevaluation of MT systems is the commonest and most widely practised in the field of MT. For this and other reasons mentioned above and since macroevaluation concerns itself with the evaluation of the quality of MT systems, it is the evaluation that is going to be used in the present study.

It is worth noticing that Van Slype (1979b:177ff) proposes three types of evaluation programme: a.) a superficial evaluation, b.) an in-depth evaluation and c.) a pinpoint evaluation (See Fig. 3.2 below). The first, which is inexpensive and easy to use, would be applied primarily at the macroevaluation level. It would be applied when each new version of an MT system becomes available. It permits an overall and comparative appreciation of the quality of each version.

The second, which is more elaborate and more expensive, would be applied primarily at the microevaluation level. Its purpose is to evaluate the acceptability and improvability of the system, and the improvements affected by simple updating of the dictionaries on the basis of the sample text. In general, it would be done on delivery of an improved version of a system of which the initial version would already have undergone one or more superficial evaluations.

The third type of evaluation would be applied on a case-by-case basis to evaluate an improvement made on a specific feature or a combination of features of the system.
MARKET STUDY

Acceptability

Intelligibility

Fidelity

Reading Time

Correction Time

Correction Rate

Improvability

MACROEVALUATION

SUPERFICIAL EVALUATION

MICROEVALUATION

DEPTH EVALUATION
3.3 The Framework Model of MTE

This section outlines the model of MTE based on the ISLE’s framework of taxonomy 3 (FEMTI, 2003) with all the requirements, criteria, metrics, evaluation types and quality characteristics relevant to the present evaluation; taking into consideration the EAGLES 7-step recipe (See, 2.5.1.3) and stages in the evaluation process (See, 2.5.1.4.3) in this study (Figs. 3.3 and 3.4).

The first step in any evaluation procedure should be to analyze user requirements, from which relevant quality characteristics are deduced and broken down in terms of reportable attributes. In parallel to the definition of the user, the objects (e.g., MT systems) of evaluation are also defined in terms of attributes. Relevant metrics are then developed. Having defined attributes and metrics for the particular evaluation environment, the next step is to decide what methods should be applied in order to perform the measurement and to obtain values for the attributes (EAGLES, 1996: 32).

It should be noticed here that the identification and explanation of all the aspects of evaluation mentioned above is mainly based on Van Slype’s critical study of MTE (1979b), the EAGLES Report (1996), the EAGLES Report (1999) and the ISLE’s Evaluation Framework (2003) for it is comprehensive, up-to-date and accessible (FEMTI, 2003).

3.3.1 Establishing Evaluation Requirements
As it is mentioned above (See also Fig. 3.4 below), the evaluation process is now broken down into four main stages: a.) establishing evaluation requirements, b.) specifying the task model and defining users, c.) designing the evaluation and defining quality model, d.) executing the evaluation and defining evaluation requirements, and further breakdown.

STEP-1

The Evaluation Purpose

STEP-2

Task Model
Defining Users

STEP-3

Defining Quality Model

STEP-4

Defining Evaluation Requirements

STEP-5

Defining Metrics & Measurements

STEP-6

Designing the Evaluation

STEP-7

Executing the Evaluation
Establishing the evaluation requirements involves three steps discussed below (Fig. 3.5 above).

According to the *EAGLES Report* (1999:3), the purpose of the evaluation is the first concern in this process. The person commissioning the evaluation might, for example, be trying to decide between competing research proposals, or be interested in buying an MT system for use within an existing translation service, or as a component of a larger information extraction system or as a starting point for a product he intends to develop further and market himself.

In every evaluation though, there is an underlying assumption that someone will use the object being evaluated. Part of defining the purpose of the evaluation is to make explicit who that person is and what needs of his it is to fulfil. Depending on the purpose of evaluation, this person is not necessarily an end-user in the conventional sense of a person using a marketed system to achieve some task. He may be a research agency, a systems programmer, a system developer, the manager of a translation service, etc. (ibid.).

The next step is to identify the types of products to be evaluated. Depending on the purpose of the evaluation, it might be a set of specifications, a research proposal, a system component, a partially developed system, a complete system and so on. There might also be additional
constraints on the product that comes into play here: Is the MT system, for instance, to accept spoken input, or to run on some specific hardware or software platform? (ibid.)

The final step in establishing the evaluation requirements is to define and specify the quality characteristics and possible sub-characteristics of the object to be evaluated. These will decide whether it is a good object of its kind or not. Not all characteristics will be of equal importance, especially when the needs of specific user are taken into account. In this respect, Hovy (1999) maintains that “in any case, the user is free to choose anything that seems important to be evaluated.” It is stated in the EAGLES Report (1999:4) that when defining the quality model, it is usual to start with high level considerations and then to refine these down until one can arrive at attributes of the system which can be measured in some way. These attributes are the bottom line of the evaluation. The values obtained for them can be combined in different ways to reflect the relative importance of the different system characteristics to a particular user or set of users.

Regarding the present study, the purpose of evaluation involves:

(I) Evaluating the external quality characteristics of each of the MT systems under testing with respect to selected attributes (e.g., functional and non-functional). The types of evaluation required here are black-box methods, declarative and performance evaluation. The functional and non-functional requirements of the present evaluation are listed below:

A. Functional Requirements:
   1. Functions that satisfy stated or implied user needs.
   2. The presence and appropriateness of a set of functions for specified tasks.
   3. General clarity and understanding of a translation and clarity of meaning of each sentence in a translation.
   4. Measurement of the correctness and faithfulness of the information transferred from the ST to the TT.
   5. Correctness of most important terms.
   6. Grammatical correctness of a sentence.
   7. Degree to which words are correctly inflected.

B. Non-Functional Requirements
   1. The relationship between the level of performance of the MT system and the amount of resources used under stated conditions.
2. The time required in translating a particular test text.
3. The ability of an MT system to be transferred from one environment to the other.
4. The adaptation of an MT system to different specified environments.
5. The capacity of the MT system with respect to the size of the text to be displayed in its text boxes (English and Arabic).
6. The modification capacity of the system.

(II) Testing the translation external quality of each MT system in this study with respect to texts representing various domains that each of these systems is expected to translate.

(III) Comparing the results obtained from the evaluation of the three competitive MT systems under testing (comparative evaluation).

(IV) Assessing both positive and negative effects of each MT system under evaluation and its adequacy in translation concerning various text types.

It is perhaps worth noticing that other kinds of evaluation like declarative (e.g., for ‘readability’ and ‘fidelity’, See 2.9.5), black-box evaluation (See 2.9.3) and performance evaluation (See 2.9.6) will also be used in this study.

Three MT systems Golden Al-Wafi, AL-Mutarjim AL-Arabey and An-Nakel AL-Arabi will be evaluated to fulfil the needs of the end-user regarding the purpose of evaluation. These Arabic MT systems will be evaluated and compared with respect to the ‘external quality’ which will be discussed with the general quality characteristics and sub-characteristics relevant to the requirements of the present evaluation in the following sub-sections.

### 3.3.1.1 Functional and Non-Functional Requirements

In general, requirements are partitioned into ‘functional’ and ‘non-functional’. Functional requirements are associated with specific functions, tasks or behaviours the system must support, while non-functional ones are constraints on various attributes of these functions or tasks. In terms of the ISO/ISLE quality characteristics for evaluation, the functional requirements address the quality characteristic of functionality (e.g., clarity, intelligibility, fidelity, style, etc.) while the other quality characteristics are concerned with various kinds of non-functional requirements (EAGLES, 1996: 23).
Figure (3.6). Mapping Categories of Requirements to Evaluation Criteria.

The approach taken in this work considers both types of requirements. These requirements are based on the definition of each one of the quality attributes under evaluation here. For instance, one of the functional requirements related to this study is “the system should reflect the ease with which an MT can be understood.” This requirement is derived from the quality characteristic of ‘readability’ (See 3.3.1.3.1). So, the functional requirements in the present work are related to quality criteria of functionality, while the non-functional ones are associated with non-functional quality characteristics such as, for instance, ‘efficiency’ and ‘portability’, Figure (3.6) above.

3.3.1.2 Quality: A Wider Outlook

With the ever-growing interest in NLP systems and MT systems grows a need for developing adequate methodologies and tools for assessing their qualities. Quality is to be defined, measured and assessed with respect to the extent to which stated or implied requirements are met. In fact, quality and productivity of software are the most important issues in the software industry today.

Although MT users and MT developers are obviously interested in different kinds of evaluations, they share a common interest concerning the quality of MT. However, different parties have different ways of looking at a product from the point of view of quality. In other words, it appears that a general definition of quality to which all people: developers, users,
managers, etc., agree is not available. It has been concluded that the interpretation and view on quality is influenced by the responsibility or role of a person. So, these different views of quality rightfully exist and cannot and should not be ignored (Kusters, Van Solingen and Trienekens, 1997).

Bruderer (1978:5) indicates that quality is a relative concept, i.e., one related to a specific object. It can be apparently measured, at least in part, but it remains much more difficult to quantify abstract phenomena than concrete things. According to him, quality can be evaluated:

a.) either positively: assessment of merits, advantages, b.) negatively: assessment of deficiencies, errors and disadvantages or c.) totally: assessment of the positive and negative aspects.

The concept of the quality of a product is, in general, unambiguous and the quality has to be assessed, not in the absolute, but according to the aims of the writer of the texts to be translated and by those who decide how it is to be distributed. Moreover, the quality achieved by a human translation (HT) cannot be expected of MT or MAHT, and the latter has thus to be used for more limited aims than the former. The evaluation criteria have to be chosen according to these specific aims. Further, since translation quality cannot be measured in the absolute, on the basis of a single criteria, its assessment should combine several different criteria (Van Slype, 1979b:34-37 and 105 and FEMTI, 2003:41). Consequently, it becomes necessary to make use only of the criteria presenting a real relevance to the aims and requirements of the evaluation.

There are increasing expectations for quality, both in the consumer and professional market. It is no longer sufficient to just deliver products which have technical excellence-products also need to be easy to use and to fit in with the work practices and activities of the consumer and professional user. Traditional approaches to quality put emphasis on meeting the specified requirements which are primarily functional. But, attempts have been made to broaden the perception of quality from a user perspective as: functionality, reliability, usability, efficiency, maintainability and portability (See Fig. 3.7). The objective of quality from the user’s perspective is for the software to exhibit excellence in the actual conditions of use (EAGLES, 1999:38-39).

The ISO/IEC 9126 standard has been revised to include a new quality model which distinguishes between three different approaches to product quality (See 2.5.1.4.2):

1. Internal quality, which is measured by the static properties of the code, typically by inspection (such as the content of error messages, vocabulary and translation rules).
2. External quality, which is measured by the dynamic properties of the code when executed (such as response time).
3. Quality in use, which is measured by the extent to which the software meets the needs of the user in the working environment (such as productivity).

![Software Quality Diagram]

External quality is a result of the combined behaviour of the software and the computer system, while quality in use is the effectiveness, productivity and satisfaction of the user when carrying out representative tasks in a realistic working environment. External measures can be used to validate the internal quality of software. Also, appropriate internal attributes of the software are a pre-requisite for achieving the required external behaviour, and appropriate external behaviour is a pre-requisite for achieving quality in use (See Figs.3.8,3.9 and 3.10) (EAGLES, 1999:38-39).

Quality in use is the combined effect of the system quality characteristics for the end-user. A product meets the requirements of a particular user if it enables the user to be effective, productive in use of time and resources, and satisfied, regardless of the specific attributes the product possesses (ibid : 40).
Since the external quality of a software product is specified in terms of the behaviour of a complete hardware/software system of which the software product is a part, it is this level of quality which fits more the non-functional requirements of the present evaluation. Here, external metrics relate to the behaviour of the software product when it is executed (e.g., time of translation, adaptability, storage, etc.). The values of external measures necessarily depend on more than the software, so the software has to be evaluated as part of a working system. It is worth noticing that the software which performs satisfactorily in one environment may show quality defects in another environment. External evaluation of quality characteristics should therefore take place under conditions which emulate as closely as possible the expected conditions of use. However, internal quality of the system is
also to be measured in terms of the static (i.e., functional) properties of the software (e.g., accuracy, suitability and wellformedness, etc.).

In fact, the software quality characteristics in the revision of ISO/IES 9126 have been redefined in terms of the capability of the software, to enable them to be interpreted as either an internal or an external perspective. The definition also refers to use under specified conditions to make it clear that quality is not an absolute property, but depends on the context of use (ibid : 39).

It is worth mentioning that the evaluation task in the present work is not complete, since a software product is typically specified by a combination of internal, external and quality in use measurements. This cannot be covered within the limits of time and space allowed and would be left for further research.

3.3.1.3 Evaluation Criteria

The actual requirements (stated or implied) play a central role in the evaluation process. These will eventually be expressed in terms of criteria (or attributes) which satisfy the following conditions (EAGLES, 1996:20-21):

1. The attributes should be adequate for expressing all explicit or implicit user requirements.
2. Their values should be obtainable by observation, by direct measurement or by deriving them from the values of other attributes.
3. The attributes should be general enough to be applied to sets of systems with similar tasks and to different classes of users or usages of those systems.

Aiming to analyze MT systems with respect to the user needs, the researcher of the present work has focused on the external quality characteristics taking the ISLE classification as a starting point for this evaluation (See 2.5.1.5.3 and Appendix A).
Ideally, an evaluation of MT systems quality should cover all the different criteria liable to be considered in a system’s working and translation. However, this is too complex a task to be done in this study.

Thus, only certain criteria from the ISLE framework have been chosen which serve the aims and requirements of the present evaluation. These are selected on the basis of their measurability too, as scoring and measurement schemes for these attributes are available from the ISLE evaluation project. So, the following external quality characteristics are the focus of evaluation.
in the present work (See Fig. 3.11 in Appendix A). The identification and explanation of each attribute in this evaluation model is mainly based on the *EAGLES Report* (1996:59-62), the *EAGLES Report* (1999:13-14) and the ISLE Taxonomy 3 (*FEMTI*, 2003: 41-57).

### 3.3.1.3.1 Functional Criteria

(A) **Functionality:**

The capacity of the software product to provide functions which meet stated and implied needs when the software is used under specified conditions. This characteristic is concerned with what the software does to fulfil needs whereas the other characteristics are mainly concerned with where and how it fulfils needs (*ISO/IEC 9126*:2001, 6.1 and *FEMTI*, 2003:41).

(1) **Suitability**

This refers to the capacity of the software to provide an appropriate set of functions for specified tasks and user objectives (*ISO/IEC 9126*:2001, 6.1.1 and *FEMTI*, ibid.)

(i) **Readability** *(or: Fluency, Intelligibility and Clarity)*

This criterion has also been called ‘fluency’, ‘intelligibility’ and ‘clarity’ which refers to the extent to which a sentence reads naturally. It is one of the qualities of the translation that can be evaluated solely on the basis of the output of the system in the TL. This quality has been merged with ‘clarity’, which was a separate criterion in earlier versions of the ISLE evaluation taxonomy. ‘Readability’ is intended to be a metric applied at the sentence-level (*FEMTI*, ibid : 42-43).

In their framework of MTE, Vanni and Miller (2001) have merged the attributes of ‘comprehensibility’, ‘readability’, ‘style’ and ‘clarity’ in the ISLE framework into a single evaluation criterion which they label “clarity”.

It is an alternative to ‘intelligibility’ (Sinaiko, 1978:7). This reflects the degree to which a complete translation can be understood. Here, it is sufficient that some clear meaning is expressed by the sentence and not that meaning of the output text reflects the meaning of the input text. Thus, no reference to the ST or reference translation is permitted. Likewise, for this measure, the sentence needs neither “to make sense” in the context of the rest of the text nor
be grammatically well-formed, since these features of the text would be measured by the ‘coherence’, and ‘syntax’ test, respectively.

Arnold (1994:275) states that the ‘intelligibility’ of a translated sentence is affected by grammatical errors, mistranslations and untranslated words. Bennett et al. (1986:95) indicate that ‘intelligibility’ is the commonest criterion in MTE and the best method of its evaluation is to ask assessors to place the TT (or each sentence of it) on a scale of ‘intelligibility’. Here, the evaluators do not have access to the ST and so do not need to know the SL. Finally, it is worth noticing that the measure of ‘intelligibility’ can at best be carried out before the ‘fidelity’ of the document has been established (Arnold, 1994:278).

(2) **Accuracy**

This criterion indicates the capacity of the software product to provide the right or agreed results or effects with the needed degree of precision. ‘Accuracy’ and its sub-qualities (i.e., ‘fidelity’, ‘consistency’ and ‘terminology’) are established by reference to the ST (ISO/IEC 9126, 2001: 6.1.2 and FEMTI, 2003: 48).

The researcher, in this study, is going to concentrate on the classical criteria of ‘fidelity’ and ‘intelligibility’ of the whole text as these are of major importance in the evaluation of translation quality.

(i) **Fidelity**

By measuring intelligibility, only a partial view of translation quality is obtained. A highly intelligible output sentence needs not be a correct translation of the source sentence. It is important to check whether the meaning of the SL sentence is presented in the translation. This property is called ‘fidelity’. Scoring for fidelity is normally done in combination with scoring for intelligibility to give the overall accuracy of translation quality. This is also achieved when the scorings of other attributes are involved (Arnold, 1994 : 278).

According to Halliday and Briss (1978:6), ‘fidelity’ (also ‘correctness’ or ‘precision’) refers to the measurement of the correctness of the information transferred from the SL to the TL. In other words, it refers to how precisely the TT conveys the meaning of the ST. Van Slype (1979b:72)
and *FEMTI* (2003:48) indicate that ‘fidelity’ is subjective evaluation of the measure in which the information contained in the sentence of the original text reappears without distortion in the translation.

(ii) Terminology
This is a subjective evaluation of the degree to how correctly the most important technical (domain-specific) terms are translated (*FEMTI*, ibid:50).

In many domains, especially for commercial translation (which is often very ‘domain-oriented’ and makes heavy use of MT), terminology translation can be extremely important (Hovy and King, 2001:23).

Since, in the present work, there are three English/Arabic MT systems with common and different domains, the evaluation of this criterion is one of the major requirements of quality assessment of each of these systems. These systems will be evaluated in terms of how correctly important domain terms are translated and how one system more correctly translates the terms of a shared domain.

(3) Wellformedness
This points out to the degree to which the output respects the reference rules of the TL at the specified linguistic level. The ISLE Project includes, here, only the four most critical categories of error typically made by MT systems involving: a.) punctuation, b.) lexical choice, c.) syntax and d.) morphology. Concentration in this work will be on the last two sub-quality characteristics (*FEMTI*, 2003:50-51).

(i) Syntax
A very short definition is given in the ISLE framework of MTE to this criterion as the degree to which the output respects the reference grammatical rules of the TL (ibid:51).

Maier, Clarke and Stadler (2001) point out that grammatical correctness refers to the recognition and correct translation of grammatical structures. Relevant to this, Nyberg, Mitamura and Carbonell (2001) indicate that the grammatical structure of each target sentence should be completely correct (i.e., no grammatical errors).

(ii) Morphology
This refers to the degree to which the output respects the reference (inflectional) morphological rules of the TL. Inflections typically carry information about number, gender, case, aspect, etc.). This quality is especially important for highly inflected languages like Arabic (*FEMTI*, 2003:52).
It should be noted that sometimes it is difficult to separate purely morphological effects from those that have their roots in syntax.

3.3.1.3.2 Non-Functional Criteria

(A) Efficiency
A set of attributes that bear on the relationship between the level of performance of the software and the amount of resources used, under stated conditions. Resources may include other software products, hardware facilities, materials (e.g., print paper, floppy disks) and services of operating, maintaining or sustaining staff (EAGLES, 1996: 60 and ISO/IEC 9126, 2001: 6.4).

Two sub-characteristics to this criterion are given in the ISLE framework (FEMTI, 2003:55): a) time behaviour and b.) resource utilization. Relevant to the requirements of the present work, only the first attribute will be evaluated involving, as sub-attributes, the input-to-output translation speed, pre-processing time and post-processing time (correction time). The second and third sub-attributes within ‘time behaviour’ are out of the scope of evaluation in this study.

(1) Time Behaviour
The definition given to this criterion in the ISLE framework (ibid.) refers to attributes of software that bear on response and processing times and throughput rates in performing its function (See ISO/IEC 9126, 2001: 6.4.1).

(i) Production Time / Input-to-output Translation Speed
According to the ISLE framework (2003:55), this is relevant to the time required for the translation of a particular test text. The importance of this parameter is different for different types of translation work. This characteristic can be divided into the following sub-characteristics:

a.) Production time for complete translation.
b.) Production time for preliminary translation.

Calculation of the production time for complete translation of each sample text and each MT system in the present work is one of the requirements of quality assessment.

(B) Portability
A set of attributes that bear on the ability of software to be transferred from one environment to another (EAGLES, 1996:60 and ISO/IEC 9126, 2001:6.6). The environment may include organizational, hardware or software environment (FEMTI, 2003:57). There are five sub-characteristics to ‘portability’. Only ‘adaptability’ will be tested in this work (See Appendix A).
(1) Adaptability
According to the EAGLES Report (1996) and ISLE framework (FEMTI, 2003:57), this refers to attributes of software that bear on the opportunity for its adaptation to different specified environments without applying other actions or means than those provided for this purpose for the software considered (ISO/IEC 9126, 2001: 6.6.1).

(C) Storage
This attribute is added by the researcher which is not found in the ISLE’s model (FEMTI, 2003) of the present evaluation. It refers to the capacity of the MT system with respect to the largest size in Kb of the text that can be displayed in the English and Arabic text boxes. This is to see whether the system runs adequately and the minimum hardware requirements it needs.

(D) Flexibility (or Extensibility)
Like ‘storage’, this is an attribute added to the present evaluation model (See Fig. 3.11 above) which concerns the modification capacity of the MT system with respect to the quantities and qualities of additions (e.g., new lexical items to its dictionary or making new dictionaries) or any other similar processes or changes.

3.3.2 Specifying the Evaluation
Here, the first step is to work out in concrete terms how the criteria reviewed in the previous section are to be measured. Different evaluation methods and tests are required in this respect. This may sometimes involve just looking something up, for instance the languages the MT system can deal with or the price of a commercial product. In other cases, there may be a test to be applied, for example, in finding out how long it takes to translate a page of text or to apply ‘cloze-test’ or multiple-choice test in the evaluation of ‘intelligibility’. In yet others, it may be necessary (or in some cases desirable) to call upon the exercise of human judgement, for example, in questionnaires, checklists and interviews to measure whether the end results of the translation process are presented in a manner that most people find attractive and easy to understand. It should be noticed that whatever metric is decided upon, it should measure what it is supposed to measure (i.e., validity), and do so consistently (i.e., reliability-not giving one answer one day and a different answer the next). This can be the most delicate and time consuming part of designing an evaluation. Metrics are discussed in more detail in the following sub-sections (See also EAGLES, 1999: 4).
Once how value will be obtained has been established, the next step is to decide on what counts as a good, bad or neutral value, and, in particular, what a satisfactory value will be for this evaluation (ibid.).

Again, several sets of criteria are considered for possible implementation in this study; it is the researcher’s aim that the measure finally chosen be as objective as possible, and thus replicable. In other words, the researcher has the goals of finding objective, consistent tests which accurately correlate with a desired goal of MTE. The ISLE Classification Framework for MTE, which concerns the quality characteristics of software products assumes the need of a quantitative evaluation leading to definition of metrics. However, the classification is not “fine-grained” enough to evaluate the quality of machine translated texts as Marrafa and Ribeiro (2001) believe. This is primarily due to the subjectivity of judgement regarding certain tests of attributes suggested in the framework.

With this in mind, metrics in the present evaluation are developed partly by consulting the taxonomy, partly by consulting the literature and partly by developing original metrics where possible. This is to make the application of metrics and results less subjective.

However, one of the ways to reduce the level of subjectivity of metrics and evaluation is utility of automatic measures. As Dabbadie et al., (2002) confirm, they provide cheap, quick, repeatable and objective evaluations. But, these cannot be applied in the evaluation of all criteria. Given the difficulty of the task, most of the quality judgements and metrics in the present work should be carried out by human evaluators. On the other hand, the number of evaluators can be a crucial factor in the objectivity of results (Koh et al., 2001). For instance, in a comprehension test, which is mainly based on subjective judgment, many evaluators seem to be needed to obtain reliable results. In this respect, Darwin (2001) states that two evaluators scored the ‘intelligibility’ and ‘accuracy’ of the test sentences in his study. Due to the limits of space and time, only two evaluators, i.e., the researcher and an experienced translator will perform the evaluation task in the present work. Further, as far as practical application of metrics and tests are concerned, subjectivity cannot be excluded—as is the case in any MTE. It can, then, be reduced. A scale of the quality of translations should be reliable, valid, objective and easy to use (Miller and Beebe, 1956:73). However, Hovy and King (2001:30) and Guessoum (2002 c and d) maintain that there is still a problem of deciding whether a 3-point scale is more effective than a 5-point scale or a 7-point scale in the evaluation processes of different criteria. In this respect, the researcher thinks that the less number of points are there on a scale, the more objective the tests and results will be. This is also confirmed by Guessoum (2002 e) who
maintains that “ideally, the smaller number of choices, the better and more objective the evaluation will be.”

In short, the purpose of this stage, i.e., specifying the evaluation is to prepare the basis for evaluation. It consists of three components (EAGLES, 1996: 147-48):

1. **Quality metrics selection**: This indicates the manner in which quality characteristics have been defined and does not allow their direct measurement. The need exists to establish metrics that correlate with the characteristics of the software product. Every quantifiable feature of software and every quantifiable interaction of software with its environment that correlates with a characteristic can be established as a metric. Metrics can differ depending on the environment and the phases of the development process in which they are used.

2. **Rating levels definition**: Quantifiable features can be measured quantitatively using quality metrics. The measured value must be interpreted as a rated value, i.e., divided into ranges corresponding to the different degrees of satisfaction of the requirements. Since quality refers to given needs, no general levels for rating are possible. They must be defined for each specific evaluation.

3. **Assessment criteria definition**: To assess the quality of the product, the results of the evaluation of the different characteristics must be summarized. The evaluator has to prepare a procedure for this using, for instance, decision tables or weighted averages. The procedure usually will include other aspects such as time and cost that contribute to the assessment of quality of a software product in a particular environment.

### 3.3.2.1 Metrics

#### 3.3.2.1.1 An Introduction to Metrics

Many people believe that translation by its nature, being subjective, cannot be measured. Others believe that almost everything can be measured, including translation. Woyde (2001:37) thinks that although there is subjectiveness to translation that cannot be measured, there are key translation qualities that can be measured and should be measured. Of course, this includes the evaluation of TM, MT and database driven publishing environments in addition to HT.
MTE has lacked standards for metrics since the beginning of MT technology, despite concrete efforts throughout its history (See also 2.5.1.5.4). Very little is said in ISO/IEC 9126 about metrics, except that the state of the art is not sufficiently advanced for standardization work to be carried out, and that only a few generally accepted metrics exist for the quality characteristics given. However, it is worth noticing that in this standard, a ‘metric’ is by definition “a quantitative scale and method which can be used to determine the value a feature takes for a specific software product.” (ISO/IEC 9126, 1991 : 1, 3.14)

This is in contradiction to the EAGLES (1996) proposal, where metrics are classified by the type of value they may take into: a.) facts, b.) judgments and c.) tests. Facts need not involve any kind of scale and judgements are qualitative rather than quantitative, since they involve the exercise of human judgement (e.g., comparative values).

According to the EAGLES (1999:46), metrics are an essential concomitant of the quality model laying at the base of any evaluation. In ISO terms, quality characteristics are broken down into sub-characteristics, which in turn are broken down into measurable attributes. Looked at from the bottom up, starting from the metrics and leading upwards towards the top level quality characteristics, it is possible to think of the product as being specified by the metrics that will be applied to it.

In fact, metrics have moved into the foreground as an area of interest. A distinction is made between metrics relating to internal characteristics (internal metrics) of the software and metrics relating to the behaviour of the software as seen from the outside (external metrics). Internal metrics are therefore measuring characteristics of the software itself such as the number of lines of code. External metrics, on the other hand, relate to the behaviour of the software when it is run. Recently, a work item on metrics for quality in use has been added (ibid.). Of course, external metrics are associated with the external quality characteristics which are the subject of evaluation in the present work. Here, applying the metric obtains a measure which indicates how the object (i.e., an MT system) being evaluated is assessed with respect to that attribute. The values obtained can be combined in different ways at higher nodes to give measures also at the higher nodes of the attribute structure.

To summarize, a metric is a measure and a method for obtaining a value for that measure. A metric is validated by ensuring that the value obtained correlates with an attribute of the quality model. Definition and validation of metrics for the quality models related with different application softwares is an area where much remains to be done.
It is worth mentioning that metrics in this study are basically quantitative (See ISO/ISLE documents) and the values are expressed in quantitative terms (e.g., ranging over the scale ‘brief’, ‘acceptable’, ‘long’, ‘too long’, etc.). These are mainly subjective measures based on human evaluation and are therefore distinguished from metrics based on objective testing. These are the most commonly used metrics in the evaluation of external quality criteria of an MT system (*FEMTI*, 2003). However, to obtain more objective values, the evaluation results of the two evaluators in the present work must coincide to a large extent.

### 3.3.2.1.2 Validity of Measures

It has already been pointed out that the validity of measures is a critical issue, i.e., if the measures used are not valid, the evaluation as a whole is worthless.

In general, both measures and the methods used to obtain a measurement must be valid, that is, a metric should measure what it is supposed to measure.

Although there are several conceptions of validity to be found in the literature, they all essentially fall under one of two broad categories: a.) internal (or contents) validity and b.) external (or criteria based) validity. Internal validity is achieved by making sure that each metric adequately measures an appropriate attribute of the object to be evaluated. This is assessed by the judgement of experts. External validity is determined by calculating the coefficient of correlation between the results obtained from applying the metric and some external criteria (*EAGLES*, 1996: 14, 29-30).

Ensuring internal validity is, in general, rather more difficult to achieve. It relies on the judgement of experts, who design the evaluation and can only be justified after the event in the light of feedback from the customers of evaluation or from other interested outsiders (ibid: 30).

External validity, on the other hand, rarely worked out formally (e.g., by actually calculating a coefficient of correlation) in the case of evaluation design, but is so often used informally to justify the choice of a measure (ibid.).

The present study relies on internal validity to determine whether the metric selected for the evaluation of each attribute is or is not well-chosen. This is based on the judgement of experts in the field of translation and computer science, in addition to the evaluators (See Appendix B).

### 3.3.2.1.3 Reliability of Measures

Measures must also be reliable if there exists a reliable method for achieving the measurement. A metric is reliable in as much as it constantly provides the same results when applied to the
same phenomena. Reliability can be determined by calculating the coefficient of correlation between the results obtained on two occurrences of applying the metric (EAGLES, 1996: 14, 30).

Regardless of the subjectivity of measures in the present study, the reliability of metrics will be found out by statistically testing and comparing the results of the two evaluators (See the next chapter) as these should correlate to a certain degree. The same results indicate the reliability of a measure if done at different times and places or if done by different people (ibid.).

3.3.2.2 Methods for System Measurement

As it has been mentioned earlier in this study, after the identification of the attributes and metrics for the particular evaluation, the next step is to decide what methods should be applied in order to perform the measurement and to obtain values for the attributes.

There is, in fact, little consensus concerning the exact nature of tests and judgements that are appropriate for software evaluation. Much has been said about the importance of software quality assurance from the viewpoint of project management and developers, but little has been presented from the viewpoint of the end-user, who so far has played a rather passive role in software development and evaluation. Moreover, the principal aim of existing testing methods as reported in the literature is to uncover software errors and not to allow the adequacy of systems for a particular user environment to be assessed naturally enough (ibid : 32).

The method of evaluating each of the attributes chosen in this work will be identified and explained (along with the metric appropriate for this purpose) in the next section. These methods are based on general software engineering principles and are user-oriented, i.e., the evaluation is seen as being performed on behalf of a user or of a class of users. Mostly, ‘black-box’ testing (where only input and output behaviour can be observed) will be applied. In addition, declarative evaluation for ‘readability’ and ‘fidelity’ (See 2.9.5), performance evaluation (See 2.9.7) for testing other criteria (e.g., speed of translation, adaptability, etc.), and adequacy and comparative evaluations (See 2.9.6 and 2.9.10, respectively) will be used in this study to see whether any of the systems in a given application domain is adequate for its particular task and the extent to which it is more suitable than the others.
Finally, the comparison will reveal which one of the systems under evaluation is the best with respect to certain tasks and purposes.

In general, methods are the way in which results are obtained. Three principal motivations may roughly be distinguished in the present evaluation with respect to methodological attempts:

1. To assess the appropriateness of a piece of software for every-day work;
2. To examine the behaviour of software under specific conditions and
3. To check the actual functionality of a piece of software.

Of course, the actual testing procedures will differ largely from one attribute to the other, but sometimes certain attributes share the same method (e.g., a 3 or 4-point scale with similar procedures in application). In most cases, MT systems are often evaluated by judgements using a rating scale (

It is worth mentioning that there are basically three types of evaluation methods: a.) inspection, b.) test and c.) judgement. The first two are the preferred methods since they are objective. Inspection can be used in looking up language covered, price and operating system. Tests can be automatic, semi-automatic, questionnaire and think-aloud protocols. Judgements which are subjective cannot always be avoided, but should be kept to a minimum. They can be questionnaire, think-aloud protocols and value on a rating scale (which is largely adopted in the present study) (ibid.).

In brief, within the framework of the present evaluation, the theoretical groundings of systems are not considered. It is mainly a black-box and comparative evaluation of three Arabic MT systems with no accesses to their inner workings.

3.3.2.3 Description of Metrics and Criteria Scoring Methods in the Present Evaluation

MT is evaluated for a number of different reasons and, when possible, these should be kept clear and separate, as different types of evaluation are best suited to measure different aspects of an MT system (See 3.2.1).

However, the evaluation methodologies for MT systems in this work will center on ‘black-box’ approaches, where global properties of the system will be evaluated, such as semantic fidelity of the translation or intelligibility of the TL output.
As it has been mentioned in the previous chapter, the ISLE Framework for the Evaluation of MT \textit{(FEMTI,2003)} provides and proposes metrics and methods for the evaluation of the characteristics of MT systems. Moreover, the criteria chosen for the present evaluation are very important in a general evaluation of MT quality. Despite the fact that the metrics proposed in the ISLE classification do not provide a sufficiently objective evaluation of these criteria, the aim in this work is to take the suggested metrics of the ISLE framework as a starting point for the present evaluation.

As for the methods of evaluation, these will be partly selected from the literature and partly developed by the researcher.

Ideally, an evaluation of MT system quality should cover all the different parameters liable to be considered in a translation. However, this is a too complex task to be done in this study. Thus, it has been decided to focus on the sentence level.

It is worth noticing that an attribute may be split into sub-attributes and the measure for the attribute (e.g., accuracy) is then composed of the measures for the sub-attributes (e.g., fidelity and terminology) according to a well defined algorithm.

The sub-sections below present a detailed identification and explanation of the human-based metrics and methods of evaluating each of the quality characteristics selected in this study.

\textbf{3.3.2.3.1 Evaluation of Readability}

According to the \textit{FEMTI} (2003:42), ‘readability’ (or fluency, intelligibility and clarity) is a sub-attribute of measuring the ‘suitability’ of a software which is a functional criterion. In the model of the present evaluation ,only ‘readability’ has been chosen among the other criteria (i.e., comprehensibility, coherence, cohesion, coverage and style) to measure ‘suitability’.

Vanni and Miller (2001) and Dabbadie et al. (2002) point out that the ‘clarity’, i.e., ‘readability’ score for a sentence is basically a snap judgement of the degree to which some discernible meaning is conveyed by a sentence. Thus, the readability/clarity score for the entire text is the mean sentence ‘clarity’ score.

It is not surprising that short sentences are found to yield artificially high clarity scores since the phenomena which make sentences longer, such as embedded sentences, quotations, and
relative clauses tend to complicate structure and necessitate a higher quality of translation to ensure clarity. For complex sentences to score well on this feature, the relationship between sentence parts has to be explicit (Vanni and Miller, 2001:3 and Dabbadie et al., 2002).

**How to Measure**

As suggested by Van Slype (1979b:67-68); Vanni and Miller (2001:3); Dabbadie et al., (2002) and *FEMTI* (2003:42), ‘readability’ can best be measured by rating sentences on a 4-point scale. This is a subjective measure which is arrived at by assigning a score between ‘0’ (meaning of sentence is not apparent, even after some reflection) to ‘3’ (meaning of sentence is perfectly clear on first reading). Here, no reference to the ST or reference translation a human translation (HT) is permitted. In addition, it has been discovered that considering each sentence in isolation and independent of the discourse structure is an important element of the test design.

In this work, evaluators will be tasked with assigning a ‘clarity’ score to each sentence in a text according to the following (Dabbadie et al., 2002 and *FEMTI*, 2003:42):

<table>
<thead>
<tr>
<th>Score</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Meaning of sentence is perfectly clear on first reading (=all the meaning);</td>
</tr>
<tr>
<td>2</td>
<td>Meaning of sentence is clear only after some reflection (=more than 50% of the meaning);</td>
</tr>
<tr>
<td>1</td>
<td>Some meaning can be gleaned from the sentence with some effort (=less than 50% of the meaning);</td>
</tr>
<tr>
<td>0</td>
<td>Meaning of sentence is not apparent, even after some reflection (=no meaning).</td>
</tr>
</tbody>
</table>

By ‘some reflection’ is meant to refer to the context or reading the text several times.

**Method**

To apply the above metric of ‘readability’, the following steps are suggested:

1. Take only the TT (the output) produced by each MT system under evaluation.
2. Detach and number the sentences of the TT from their context.
3. Read each MT sentence.
4. Rate each MT sentence on a 4-point scale described above.
5. Obtain the total score of ‘readability’ per text by summation of sentence scores.

6. Obtain the percentage value of ‘readability’ (i.e., calculate the average of the sentence ratings per text) depending on the maximum score on the scale.

3.3.2.3.2 Evaluation of Fidelity

The evaluation of the fidelity of a text involves testing the degree to which the information contained in a sentence in the original text has been reproduced without distortion in the translation, i.e., syntactic and semantic fidelity. Moreover, ‘intelligibility’ (or ‘readability’) and ‘fidelity’ are the principal attributes that declarative evaluation measures.

According to White (in Somers, in press), in a fidelity evaluation, concentration is on “how well the content of the ST is conveyed in the TT.” Thus, measuring fidelity requires an awareness of the source document, which usually means a translator is required to make the judgements. To achieve simplicity and portability, judgements are kept monolingual if possible. So, two versions of the fidelity measure have been developed. One involves rates with expertise of the SL (i.e., directly comparing the informativeness of the Arabic and the original) and the other where the raters are monolingual (i.e., comparing the MT output with expert translation(s)). Obviously, scorers have to use the second method in case they cannot speak the SL.

Further, Darwin (2001) indicates that in measuring ‘fidelity’, the higher the amount of information in the translated text not found in the original text, the lower the level of fidelity. Additionally, omissions of actual word or sentence segments from source to target should also be checked.

How to Measure

According to the ISLE framework (FEMTI, 2003: 48) which is based on the measurements presented by Van Slype (1979b: 72-78), fidelity can be measured by any of the metrics given below:

(1) Carroll: Rating of sentences read out of context on a 9-point scale.
(2) Crook and Bishop: Rating on a 25-point scale.
(3) Halliday: Assessment of the correctness of the information transferred.
(4) Leavitt: Rating of text units read on a 9-point scale.
(5) Miller and Beebe–Center: Rating of a text on a 100-point scale.
Miller and Beebe-Center: Shannon measurement of the quality of information transferred.

Sinaiko: Re-translation.

Van Slynke (in Van Slynke’s Critical Report) : Rating of sentences read on a 4-point scale.

White and O’Connell : DARPA 94 measure.

It is the 4-point scale which is based on Van Slynke (1979b:28), Nagao as described in Jordan, Dorr and Benoit (1993:55) and Miller et al., (2001) that is confirmed to be the best to apply and resulted in the highest inter-rater consistency. In this respect, Hovy (2002d) maintains that psychology research has shown that scales up to 5 points are reasonable for people. More are confusing. Thus, ‘fidelity’ can best be measured by the following 4-point scale which is adopted in the present evaluation:

<table>
<thead>
<tr>
<th>Score</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Completely or almost completely faithful (=all the information);</td>
</tr>
<tr>
<td>2</td>
<td>Fairly faithful (=more than 50% of the original information passes in the translation);</td>
</tr>
<tr>
<td>1</td>
<td>Barely faithful (= less than 50% of the original information passes in the translation);</td>
</tr>
<tr>
<td>0</td>
<td>Completely or almost completely unfaithful (=no information).</td>
</tr>
</tbody>
</table>

It is worth noticing that the ‘fidelity’ rating should, generally, be equal to or lower than the ‘intelligibility’ rating, since the unintelligible part of the message is of course not found in the translation. Any variation between the ‘intelligibility’ rating and the ‘fidelity’ rating is due to additional distortion of the information, which can arise from:

1. Loss of information (silence)- example: word not translated.
2. Interference (noise)- example: word added by the system.
3. Distortion from a combination of loss and interference–example: word badly translated.

Detailed analysis of the ‘fidelity’ of a translation is very difficult to carry out, since each sentence conveys not a single item of information or a series of elementary items of information, but rather a portion of message or a series of complex messages whose relative importance in the sentence is not easy to appreciate (FEMTI , 2003 : 49).
Method
In this study, ‘fidelity’ scores will be computed in the following manner (since the evaluators are aware of the SL, it is necessary to perform this test by judging the ‘fidelity’ of the translation with respect to the original text):

1. Both the ST (the original text ‘input’) and the TT (the MT output) should be segmented, i.e., with each text separated into sentence fragments to appear next to the corresponding fragment in the other text (bilingual variants).
2. Once each MT sentence is lined up with its equivalent in the ST (given the same number of its equivalent in the ST too), evaluators grade each (after being read and compared with the original sentence) on the 4-point scale mentioned above.
3. Then, the total score of ‘fidelity’ per text is obtained by summation of each sentence score.
4. The last step is to calculate the average of the sentence ratings per text, i.e., the percentage value of ‘fidelity’ depending on the maximum score on the scale.

3.3.2.3.3 Evaluation of Terminology
Vanni and Miller (2001) developed an MT evaluation measure based on the percentage of domain-specific words from the ST that were correctly rendered in the translation.

The important consideration for this test is that domain terms be exact. Therefore, variations counted wrong for the purposes of this test include these stemming from the occurrence of forms irrelevant to the TL, not-translated forms, synonym usage, misspellings, wrong ordering for phrasal constituents of terms, errors of category at form, etc. Raters have to mark anything other than an exact match as wrong even though the variations encountered may be correct from the language perspective (ibid.).

These are important issues for an algorithm designed to characterize a system’s terminology-handling capability. Because the guidelines for this test are precise, evaluators can be strict in their implementation of them.

How to Measure
The metric is the percentage of correctly translated domain terms (Miller et al., 2001) or the percentage of other input/output domain terms mistranslated (FEMTI, 2003:48). In other words, it involves the ratio of the number of domain terms appearing correctly in the translation (i.e.,
MT output) to the total number of domain terms in the HT or it deals with the ratio of the number of domain terms mistranslated in the TT to the total number of domain terms in the ST (input).

Here, the black-box evaluation is applied where we compare between two outputs (i.e., a human and a machine translated texts) or an input representing the ST and the MT output TT. The last method will be adopted in the present study.

**Method**

The method utilized in the evaluation of ‘terminology’ is not based on the identification of potential domain-identifying terms in the human translated text which determines whether they occur in the MT text. Rather, in the event that the evaluators have knowledge of the SL, an alternative methodology would be:

1. To identify domain terms in the SL document. (i.e., key terms of a given domain).
2. To identify acceptable translations for each domain term (from specific dictionaries and domain-specific texts). Here, an adequate HT can be used as a guideline merely.
3. To determine the percentage of these domain terms that will be correctly represented (or mistranslated) in the MT output.

During the test application, systems receive a point (i.e., number 1) for each term for which the translation matches the acceptable translation exactly (HT of a term), no point otherwise. The final score is the percentage of exactly-matched translations of key terms.

### 3.3.2.3.4 Evaluation of Syntax

Simply, the test aims at evaluating the grammatical correctness of the sentence.

**How to Measure**

Two measurements of syntax are presented here:

1. Vanni and Miller (2001) state that the syntax score is based on the minimal number of corrections necessary to render an MT output sentence grammatical.
2. The second syntactic score is suggested by Hovy and King (2001: 20-22) based on the DARPA evaluation fluency measure (White et al., 1994) involving human rating of sentences on a 5-point scale. The test measures the syntactic correctness of the given fragment (e.g., a full syntactic unit such as a sentence). Each fragment in the TT is rated on a 5-point scale, where:

<table>
<thead>
<tr>
<th>Score</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Fragment is perfectly grammatical;</td>
</tr>
<tr>
<td>4</td>
<td>Fragment is almost perfect, but has one or two small problems;</td>
</tr>
<tr>
<td>3</td>
<td>Fragment is reasonably grammatical, but has one or two major problems;</td>
</tr>
<tr>
<td>2</td>
<td>Fragment is almost ungrammatical, with serious problems;</td>
</tr>
<tr>
<td>1</td>
<td>Fragment is ungrammatical.</td>
</tr>
</tbody>
</table>

**Methods**

A.) The scoring algorithm followed in the measurement of syntax with respect to the minimal number of corrections in the MT sentence is presented below:

1. The MT output of each text is divided into numbered sentences.
2. Each evaluator is tasked with transforming each sentence in the MT output into a grammatical sentence by making the minimum number of replacements, corrections, movements, deletions, or additions possible.
3. These changes are then scored with the exception that corrections and replacements are counted as a single category.
4. The syntax score for each sentence is then calculated as the ratio of the number of changes for each sentence to the number of words in the sentence.
5. The percentage of the overall syntactic score for the text is also calculated in an analogous manner.
It should be noticed that the lower the number of grammatical corrections per text, the higher the syntax score of the text will be.

B.) The steps to be followed in the evaluation of syntax with respect to the rating of translated sentences on a 5-point scale are:

1. A reference human translation serving as an adequate translation is required here. This is placed next to each of the translations of the ST (MT), to be used as a comparison against each one.

2. Before the test is performed, both the reference HT as well as each of the MTs should be segmented, with each text separated into sentence fragments to appear next to the corresponding fragment in the translation.

3. Once each translation (MT) is lined up with its equivalent reference HT, evaluators grade each sentence on a scale of one to five, where five represents a sentence is perfectly grammatical (i.e., no corrections or replacements are made).

4. Then, the total score of ‘syntax’ per text is obtained by summation of each sentence score.

5. Finally, the percentage value of ‘syntax’ per text depending on the maximum score on the scale is calculated (i.e., the average of the sentence ratings per text).

Due to limits of time and space either one of the above-mentioned methods and measurements will be adopted in the present evaluation.

3.3.2.3.5 Evaluation of Morphology

As mentioned in the previous pages, evaluation of morphology refers to the degree to which words are correctly inflected.

How to Measure
According to Vanni and Miller (2001), the morphological score is calculated as the number of morphological corrections to the MT output divided by the total number of inflectable words in the output text.

**Method**
The following steps are used in the evaluation of the criterion of morphology:

The inflectable words in the MT output and the HT are counted and underlined.

Each evaluator is tasked with making all the morphological corrections in the output text making use of the HT as a guideline.

The morphological score of each sample text is expressed in the form of percentage ratio, i.e., the number of morphological corrections to the inflectable words in MT output.

What is important, here, is that the less the number of morphological corrections, the higher the morphological score of the whole text will be and quality gets better.

**3.3.2.3.6 Evaluation of Input-to-Output Translation**

Production time is an absolute evaluation value resulting from the number of words processed per hour written in a catalogue and the benchmark test. It is a performance evaluation of the speed of translation, i.e., the translation production time. This involves the time between a request for a translation and reception thereof (JEIDA, 1994:114) and (FEMTI, 2003:55).

**How to Measure**

Here, the score is the time required for carrying out the MT text. The evaluators are going to measure the time of translation of each text with
respect to special-domains. The speed of translation is measured by calculating the time consumed in translating a certain number of words of a particular text. This could be expressed in either the number of words per time unit or the time consumed per 100 words (Van Slype, 1979a:88). Measuring the speed of translation, certain things should be taken into consideration such as: a.) The typical translation speed of the MT system indicated by the manufacturing company, b.) The computer’s Central Processor Unit (CPU) and cache memory and c.) The used capacity of the hard drive which is loaded by softwares.

**Method**

To measure the speed of translation for each of the three MT systems under investigation, the evaluators are required to follow the steps presented below:

1. Calculating the time which each text of the twelve selected domains takes in translation using a stop-watch.
2. Expressing the translation speed for each text using the form 
   \( \text{(number of words/time of translation)} \).
3. Finding out the average value of speed for the twelve sample texts translated by each MT system considered. This is to represent the total average speed of translation of each MT system (one value for each system).
4. Transforming the average value of speed of each MT system (obtained from the previous step) into a percentage value. This results from dividing the average value of speed of each MT system by the maximum average speed scored by any one of the three MT systems (which is considered the ratio of the translation speed to a typical speed).
3.3.2.3.7 Evaluation of Adaptability

This is to test the MT system’s adaptation to different specified environments merely, i.e., the environment in which the user uses the product (*EAGLES*, 1999:44). According to Popescu-Belis (2003 a), the word environment, here, refers to “different software and hardware with which the MT system can be installed and run”. As Hovy (2003) suggests, the environment is the “operational setting” of the system which typically includes:

a.) what machine it is running on,
b.) what other software it requires,
c.) what operating system it is run under and
d.) how much disk space and memory is available.

**How to Measure**

The evaluators will test the system’s adaptability while it is being operated. Ideally, a checklist is used for testing purposes taking up the feature that is being applied, the value and most importantly, the related quality characteristic. Thus, inspection of the three participating MT system’s adaptability is based on the evaluators’ observations during system operation. The results will be assigned on a checklist prepared for this purpose (See Appendix C) where the following rating score covering five features is used for performance rating. It is worth noting that the checklist values obtained from the evaluation of ‘adaptability’ should be calculated in terms of percentages. These percentage values are essential for the estimation of the overall software quality, as illustrated below:
<table>
<thead>
<tr>
<th>Value</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Very good / superior adaptability ( =100% )</td>
</tr>
<tr>
<td>3</td>
<td>Good adaptability ( =75% )</td>
</tr>
<tr>
<td>2</td>
<td>Fair adaptability ( =50% )</td>
</tr>
<tr>
<td>1</td>
<td>partial adaptability ( =25% )</td>
</tr>
<tr>
<td>0</td>
<td>No adaptability ( =0% )</td>
</tr>
</tbody>
</table>

Accordingly, if the MT system has a superior adaptability (i.e., it runs on any machine of whatever specifications and under any operating system and is compatible with other softwares), it will take score 4 in the performance rating which equals a percentage of 100%. Its adaptability is good (i.e., will take score 3 =75% ), if it either cannot run under one or two operating systems or it cannot be installed on machines with certain specifications. If it is fairly adaptable, i.e., it cannot run on a certain PC with certain specifications and cannot run under one of the operating systems, it will take score 2 in this rating scheme which equals a percentage of 50. It will take score 1 (=25%), i.e., poor adaptability if it has very limited adaptability with specific machines and operating systems. Finally, if it is unadaptable, it will take score 0 on this scale ( = 0%) .

**3.3.2.3.8 Evaluation of Storage**

Evaluation of ‘storage’ mainly concerns itself with testing the capacity of the MT system in terms of the largest size of the text in Kb that can be displayed in the system’s English and Arabic text boxes.

**How to Measure**

The final percentage value of this feature can be found for each MT system in this evaluation . The results will be obtained by dividing the storage value of each MT system under testing by the typical storage suggested in this
study (i.e., the highest storage in Kb that any of the three Arabic MT systems can achieve).

3.3.2.3.9 Evaluation of Flexibility

Here, the MT system’s capacity with respect to the dynamical addition of new words and dictionaries or any other similar processes will be tested by the evaluators.

How to Measure

Like the evaluation of ‘adaptability’, testing the MT system’s flexibility requires the use of a checklist where the results will be assigned. A rating score covering three features will be used for performance rating (Appendix C). This is based on the evaluators’ judgements as a result of observation and testing the system’s capabilities while it is being operated. For the representation of the evaluation results (i.e., the checklist values) in terms of percentages, the following rating scheme is adopted:

<table>
<thead>
<tr>
<th>Value</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Very good or superior flexibility (= 100%)</td>
</tr>
<tr>
<td>1</td>
<td>Limited or fair flexibility (= 50%)</td>
</tr>
<tr>
<td>0</td>
<td>No flexibility (= 0%)</td>
</tr>
</tbody>
</table>

On the basis of this, if an MT system has superior flexibility with respect to the addition of new words (to update its dictionaries) and the making of new dictionaries, it will take score 2 in the rating scheme (=100%). If it is fairly flexible, it will take score 1 (=50%) which means either new words or new dictionaries cannot be added to the system. Finally, if it is completely
inflexible (i.e., where neither new words nor new dictionaries can be added to the system), it will take a value of 0, i.e., 0%.

3.3.2.4 GSQ Weighting Scheme

Regarding the final calculation of the average value of General Software Quality (GSQ) in the present evaluation, there is a need to combine different scores into the ‘mother node’ as a general score. According to Hovy (2002c), a simple method of combining these scores is to decide in each domain, how important the sub-characteristics of a general criteria are relative to each other and whether they are equally important or one is more important than the other(s). Moreover, Popescu-Belis (2002a) confirms that there is no “fair” distribution of the weights (or percentage values of criteria) here, because these should be fixed according to the destination of the MT systems. This depends on the needs of the users or the context, i.e., it must capture the purpose of the system and how it will be used. In this study, concentration is on the functional criteria, since they represent the main function of the MT system and the major requirements of the current evaluation. They are also the most essential and difficult problem of any evaluation process. Such a weighting scheme enables the evaluators to measure the overall value of GSQ of an MT system after obtaining the results of the calculation of the total average values of each MT system in terms of the functional and non-functional criteria selected.

According to the EAGLES (1996:30), measures may be compound. In other words, the value of an attribute (i.e., criterion) may itself be structured. The value taken by an attribute may therefore be a composite, calculated on the basis of putting together an average or even a weighted average of the values assigned to two or more other attributes. Any one of these attributes
may in its turn be a composite. In other words, there is no theoretical limit on the depth of the hierarchy of attribute–value pairs. In brief, the values of the sub-attributes combine to a value for a mother node to reflect relative importance and strength of the components in a specific evaluation. Further, finding out the GSQ or the total performance of an MT system is essential in case of comparing systems (as it is the case in this study).

In the present evaluation experiment, the functional criteria are critical and of major importance. Thus, they have been assigned the highest percentage values with respect to the weighting scheme of the overall value of GSQ of an MT system. To avoid counter arguments and as the sub-characteristics under ‘Accuracy’, ‘Suitability’ and ‘wellformedness’ are closely related from the translation view point, they have been given equal percentage values in the weighting scheme of the GSQ. The non-functional sub-characteristics have been equalized too regarding their percentage values in this weighting scheme for the same reason. The ones chosen for the present evaluation, i.e., speed of translation, adaptability, storage and flexibility are relevant to the translation process.

On the basis of this, the distribution of percentage values of the functional and non-functional criteria in the weighting scheme of the overall GSQ is shown below. It is worth noticing that the total performance of an MT system, i.e., the GSQ gets a sum value of 100%. So, to make our distribution as fair as possible with respect to the relative importance of criteria (giving primacy to the functional ones), a percentage value of 80 has been assigned to the functional criteria; whereas the non-functional criteria have been given a percentage value of 20 (See Fig. 3.11 in this chapter). Thus:
1. The percentage values of the functional criteria in the weighting scheme of the overall GSQ are as follows: ‘Functionality’ = 80% where the sub-characteristic of ‘Suitability’ = 16%; while each of ‘Accuracy’ and ‘Wellformedness’ = 32%. Each of the terminal sub-characteristics under these three mother node criteria, in turn, takes the percentage value of 16% for the reasons mentioned above.

2. The percentage values of the non-functional criteria in the weighting scheme of the overall GSQ are as follows: ’Non-functional criteria’ = 20% where each of the sub-characteristics under this mother node criteria, i.e., ‘Efficiency’/‘Speed of translation’; ‘Portability’/‘Adaptability’, ‘Storage’ and ‘Flexibility’ takes a percentage value of 5%.

Hovy (2002d,e and f) confirms that “this breakdown looks fine” to him and he sees “no obvious points that anyone could complain about.” He maintains that the distribution of percentage values like the above-mentioned depends on the personal preference for relative weightings that is based on the aims and requirements of a specific evaluation (here, the requirements are related to the end-user). For any study of this type, the most important thing is to select some weighting scheme (such as this) and then also to record all the raw numbers as well. He can “see nothing obviously wrong or problematic” in this respect (this is also confirmed by Guessoum (2002 f and g). In addition, Popescu-Belis (2002 b and c) indicates that the important thing is not to have one “absolute” evaluation scheme, but to adapt this to the user needs. So, there are no “absolute” criteria, but whatever the conventional criteria are, they should be the same for everybody. Further, King, Popescue-Belis and Hovy (2003:3) maintain that the important point here is not the exact definition of any one quality characteristic or of its sub-characteristics, it is rather that MT has a multitude of potential uses in a multitude of different contexts. In any specific context, some characteristics may be important, others not, to the point where a characteristic which is a sine qua non in one context may be completely irrelevant in another. And, it is precisely because the relative importance of individual quality characteristics is never the same in two different work contexts that the MT evaluator is
tempted to feel that he is tackling a problem which has never been tackled before, and therefore, to design his evaluation from scratch.

After all, for calculating the overall GSQ, the following formula is suggested and applied by the researcher according to the distributed criteria weighting scheme:

\[
\text{GSQ}\% = 0.16R + 0.16F + 0.16T + 0.16S + 0.16M + 0.05P + 0.05A + 0.05G + 0.05X \\
\]

Where:

- \( R \) = Readability
- \( F \) = Fidelity
- \( T \) = Terminology
- \( S \) = Syntax
- \( M \) = Morphology
- \( P \) = Speed of Translation
- \( A \) = Adaptability
- \( G \) = storage
- \( X \) = Flexibility

3.3.3 Producing the Evaluation Plan

Once the metrics and methods of evaluation have been decided upon, it is possible to plan the evaluation itself. Who will produce test materials if they are required? When will any tests be administered and by whom? How are the results to be reported and communicated? All this is part of producing
the evaluation plan, which is discussed further in the next chapter of this study (See *EAGLES*, 1999 : 4).

### 3.3.4 Executing the Evaluation

According to EAGLES (ibid.), once all the preparatory work has been done, the evaluation itself can be carried out. This will involve applying the metrics that have been decided upon, determining whether the values obtained fall into the good, bad or indifferent range and whether the values and combinations of values are satisfactory. The results will be combined to give an overall assessment of the product that has been evaluated, which will tell us whether the product is likely to meet the needs of the user whose requirements were articulated at the beginning of creating the evaluation design.

This last step of the Evaluation Process Model will be the interest of the coming chapters of the present work.

### 3.4 Summary

In fact, MT evaluations are classic hot-bottom in the MT community as a source of information, argument, dispute and sometimes even misinformation. Yet, at the same time, the need for effectively evaluating what MT systems can and cannot do and how well they do them remains.

Recent work seeks to standardize, normalize and characterize MTE. The MTE taxonomy of the ISLE is an important step in this respect, where high-level needs are broken down into lower level attributes that can be tested and measured. It is this model of MTE which has been adopted in the present work.
In this chapter, two interrelated aspects of MTE have been discussed: the framework model of MTE and the important stages of the evaluation process. It has been stressed that the evaluation process model is part of the guidelines for use of the quality characteristics. The various phases of the process have been discussed in detail too including: 1.) establishing evaluation requirements, 2.) specifying the evaluation, 3.) producing the evaluation plan and 4.) executing the evaluation. The quality requirements (both functional and non-functional) have been expressed in terms of quality characteristics and sub-characteristics. Much attention has been paid to the functional requirements and criteria in the present work as they are the most common and essential criteria in every evaluation process. Moreover, they are more closely related to the researcher’s field of study and, to the researcher’s knowledge, they have not been tested and evaluated for Arabic MT systems before.

The stage of evaluation preparation in its three sub-phases: a.) quality metrics selection, b.) rating levels definition and c.) assessment criteria (i.e., methods of evaluation) definition has been given much concern in this chapter. For each criterion under evaluation, the appropriate metrics, scoring schemes, the rating levels, and the methods regarding the overall judgement of the product, have been identified and explained. Thus, on the basis of this stage which is considered the core of the evaluation process, the other remaining phases of this process, i.e., producing the evaluation plan and executing the evaluation can smoothly take place. These will be the focus of attention and discussion in the next chapters of this study.
CHAPTER FOUR
EXPERIMENTAL DESIGN
AND EXECUTION OF MT EVALUATION

This chapter describes an MTE experiment where emphasis is placed on the quality of output and extent to which it is geared to different user’s needs. The ISLE taxonomy for the evaluation of MT (FEMTI, 2003) is used as the basis for this evaluation task. The design and execution of the evaluation follow the methodology worked out by the EAGLES (1999). The evaluation, therefore, constitutes a standard test bed application of this methodology (i.e., task-oriented testing and benchmark testing). This methodology is an effort to characterize MT output quality in functional terms while responding to the established desiderata for MTE.

Designing the evaluation involves producing an evaluation plan which describes the evaluation methods and the schedule of the evaluators’ action.
However, the last step of the Evaluation Process Model (i.e., executing the evaluation), consists of three steps, namely: a.) measurement, b.) rating and c.) assessment (See 3.3.3 and 3.3.4, respectively).

Although the main emphasis is on the functionality of the MT systems being evaluated, the present evaluation intends also to target the assessment of certain computational quality characteristics and their limitations for the potential users and purchasers. In brief, the evaluation aims at investigating which of the three MT systems under consideration shows the best quality of translation and performance with respect to the various text domains it would have to translate.

The proposed model for the functional attributes is a black-box type comparative and adequacy-oriented evaluation. In other words, there is no interaction with the systems tested and the goal is to determine whether output is actually helpful to the user groups in question. As for the non-functional criteria, the evaluation model is said to be of the comparative performance and adequacy-oriented type.

In addition, the chapter aims at outlining the test plan, the description of the test data, test types and instruments and other evaluation requirements. Further, the chapter intends to describe the measurement of the selected quality characteristics and the general method of system’s quality assessment.

4.1 Key Concepts in Evaluation

In general, evaluation is a decision about the significance, value, or quality of something based on a careful study of its good and bad features (EAGLES, ibid.).
Struening and Guttentag (1975:520) indicate that the role of evaluation research differs from that of conventional research in its explicit attempt to make value judgements. It goes beyond hypothesis testing, if it is concerned with that at all. It is generally directed toward determining what is most effective, valuable, desirable, or useful, rather than simply whether or not a hypothesis was supported. It is concerned with the determination of relative value.

Evaluation and assessment are implicit aspects of any human activity. In the language technologies, as in most software development enterprises, evaluation measures the state of a particular model (e.g., a prototype or a product) in terms of the expectations it is committed to meet, the general expectations of models of that type and the place of that model among other models which are equivalent in some respect (White, 2003).

With respect to MT, evaluation projects aim at measuring usefulness, success and efficacy of software component that translates between two human languages (FEMTI, 2003:9). In addition, evaluation is seen as a way of defining de facto standards which products should meet. It is also seen as a way of helping the developer to ensure that his product would meet market needs and find market acceptability whilst at the same time providing the consumer with a way of deciding what product is best for him (King, 1998:8).

The EAGLES Evaluation Working Group (1999:18-20) describes evaluation as a function saying: “to evaluate is to determine what something is worth to somebody”. This can be summarized in the following formula:

\[ O \times U \rightarrow V \]

Where:
O is a set of objects
U is a set of users
V is a set of values

Here, ‘V’ represents the idea of utility that drives any evaluation. The basic idea is that evaluation expresses what some object is worth to somebody; ‘V’ expresses “worth”. As for ‘O’, it represents objects of evaluation. This refers to anything that can in principle be evaluated (e.g., MT system). Finally, ‘U’ represents users, i.e., people or organizations potentially interested in numbers of O. In addition, all the factors which are often called environmental or situational variables help to define the user’s desires, and are therefore part of U. This also involves the constraints imposed by sub-components of an overall system which might fulfil the user needs.

In the present evaluation experiment, concentration is mainly on the quality of the three Arabic MT systems. It is worth noticing that the ISLE Evaluation Working Group (FEMTI, 2003:38) maintains that the quality of the translation can be evaluated in two modes: a.) without adjustment, i.e., this aims to evaluate the quality of translation before the dictionary (e.g., user dictionary) and/or grammar is adjusted. This is also an absolute evaluation of the system, and b.) With adjustment, i.e., this aims to evaluate the quality of translation after the dictionary and/or grammar is adjusted. The higher quality of translation the user needs, the more severely the evaluation is made. In this respect, this item shows the degree of the user’s satisfaction with the system. It is this second view that is adopted in this study to judge the MT systems under consideration in terms of all their potentialities.
4.2 Experimental Design

One point to notice is that evaluation design is seen as being specific to each programme, i.e., there is no general framework which could serve as guideline for all evaluation designers.

In the present experiment, the evaluation framework deals with the selection of materials, MT systems tested, evaluators, MT system users, test types and instruments used and a view on quality characteristics, metrics and end results (See Fig. 4.1).

4.2.1 Text Sampling

The first stage of the experiment is the selection of materials to support the testing of the MT systems being evaluated.

A sample of English STs has been selected for English-into-Arabic MT and evaluation. The texts chosen are representative of twelve various special-domains that are most suitable for MT, i.e., scientific, technical and commercial fields. Also, they represent most of the domains that the manufacturing companies have claimed their systems can best translate.
The SL texts have mostly been collected from several University textbooks. Also, other texts have been taken from the United Nations Development Programme (UNDP) and Baytul-Hikma documents. Most of these texts have been selected on the basis of being rich in domain-specific terminologies.

It should be noted that for each of the sample texts in this study, an Arabic human translation has also been found. These Arabic translations have been performed by university professors who are specialists in these fields of knowledge and familiar with the domain-related terminologies, in addition to English/Arabic professional translators. Nevertheless, each of these Arabic translated texts has further been revised or even retranslated by an expert translator for sentence rearrangement; accuracy of spelling, vocabulary and grammar and for supplying untranslated and/or mistranslated words and terminologies. These revised texts act as adequate translations for the sake of comparison. In this respect, Van Slype (1979b:179) maintains that in the cases where a high quality translation is required (e.g., for scientific/technical, official and publication purposes), a text translated by a human translator and revised would be needed.

In the present evaluation, three versions of the sample texts should undergo comparative treatment: a.) the original text in the SL, b.) a TL text translated by an MT system and not post-edited (i.e., raw translation) and c.) a text translated by a human translator and post-edited (i.e., an adequate translation)(1).

The SL texts and their parallel texts (i.e., MT and HT) have been aligned and segmented into equal number of sentences, i.e., a correspondence has been established between the linguistic units of analysis (e.g., sentences) of the text in the SL and the same text in the TL (i.e., MT and HT). This is not easy as it sounds. Quite often, a sentence in one text will correspond to two
or more sentences in the parallel TL text or vice versa. The order of the linguistic elements may differ from one language to another too.

The unit of translation to be analyzed and evaluated, in this work, is the entire sentence—a grammatical unit that begins with a capital letter and ends with a period (An-Nakel, 2001:20). Thus, for each English SL sentence, an Arabic (MT and HT) TL correspondent has been assigned.

On the basis of this, extracts of 20-29 sentences constituting ‘self-contained’ texts (i.e., complete paragraphs or sections) have been taken from each of the twelve text types under evaluation (See Table 4.1). Each SL sentence in the text it represents has been given a sequence number which should be the same in all its correspondent sentences (MT and HT) in the TL texts. Consequently, for each SL text, in this work, there are three MT texts and one adequate HT in Arabic, i.e., the sample of texts consists of: (12) SL texts, (36) MT texts and (12) HT texts.

So, a total of (268) English SL sentences are used as input data in testing and estimating the cognitive and linguistic properties of sentence level translations of the three Arabic translation software packages. It is worth noticing, here, that the number of sentences selected for the evaluation experiment is within the range of the evaluation standards(2). Accordingly, Van Slype (1979b:166) confirms that the sample of texts must be compiled on empirical bases:

1. Volume of text of the order of 5,000-10,000 words, i.e., 250 to 500 sentences.
2. Significant passages (5-20 sentences) selected from documents belonging to (4-6) separate categories.
4.2.2 Three Systems Tested

In the present evaluation experiment, three English-into-Arabic MT systems have been tested (See 2.11.4.8). The overall goal of this evaluation is a comparison of these systems resulting in recommendations on which system to apply for which purpose. Both AL-Mutarjim AL-Arabey 3.00 and Golden AL-Wafi 1.00 have been registered on the researcher’s personal computer

Table (4.1). The Various Sample Texts in the Present Evaluation each with the Number of Sentences Selected.

<table>
<thead>
<tr>
<th>Text Type</th>
<th>Number of Sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological</td>
<td>22</td>
</tr>
<tr>
<td>Chemical</td>
<td>24</td>
</tr>
<tr>
<td>Computational</td>
<td>20</td>
</tr>
<tr>
<td>Financial</td>
<td>26</td>
</tr>
<tr>
<td>Legal</td>
<td>20</td>
</tr>
<tr>
<td>Medical</td>
<td>22</td>
</tr>
<tr>
<td>Military</td>
<td>22</td>
</tr>
<tr>
<td>Petroleum</td>
<td>29</td>
</tr>
<tr>
<td>Physical</td>
<td>22</td>
</tr>
<tr>
<td>Social</td>
<td>20</td>
</tr>
<tr>
<td>Commercial</td>
<td>21</td>
</tr>
<tr>
<td>Water Engineering</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>268</strong></td>
</tr>
</tbody>
</table>

(PC) only through the web site of the copyright holder (ATA Software Technology Ltd.), i.e., they cannot be reinstalled on any other PC. The hardware/software features of the researcher’s PC are presented in Table 4.2 below:
Table (4.2). Hardware and Software Features of the PC Used in the Present Evaluation.

<table>
<thead>
<tr>
<th>Hardware / Software</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Processor Server</td>
<td>Pentium 3</td>
</tr>
<tr>
<td>Main Processor Speed CPU</td>
<td>730 MHz</td>
</tr>
<tr>
<td>RAM</td>
<td>128 MB</td>
</tr>
<tr>
<td>Cache Memory</td>
<td>128 K</td>
</tr>
<tr>
<td>Hard Disk Size</td>
<td>30 GB</td>
</tr>
<tr>
<td>Hard Disk Space Used</td>
<td>31%</td>
</tr>
<tr>
<td>Windows</td>
<td>98 and XP</td>
</tr>
<tr>
<td>Office</td>
<td>97 and XP</td>
</tr>
<tr>
<td>Number of Programmes Installed</td>
<td>32</td>
</tr>
<tr>
<td>Type of Programmes Installed</td>
<td>Office, Graphics, MultiMedia, Dictionaries, Camera, Games</td>
</tr>
</tbody>
</table>

An-Nakel AL-Arabi 2.00 has also been installed on the same PC, but it can be reinstalled on any other PC running Windows 95, 98 or Millennium (ME) only. In this work, both Al-Mutarjim AL-Arabey 3.00 and Golden AL-Wafi 1.00 run under Windows XP, while An-Nakel AL-Arabi runs under Windows 98.

It should be noted that the researcher has designed user dictionaries (for unknown/untranslated words and terms and for domain-related terminologies inserted and added by the researcher), which took between 3-6 hours of continuous work on the computer with each text type, for both AL-Mutarjim AL-Arabey and An-Nakel AL-Arabi. In addition, a preference dictionary for each text type in the present evaluation has also been made for certain words and terms automatically and randomly selected by AL-Mutarjim AL-Arabey MT system merely (Table 4.3).
As for the availability and test cost of each of these MT systems, they have all been purchased from the local market. Al-Mutarjim AL-Arabey 3.00 and Golden AL-Wafi 1.00 have only recently been provided for the researcher on a special order from UAE agents of ATA Software with a special discount in comparison to their real price in the global market\(^3\)\(^4\) (Table 4.4).

To attain the greatest possible objectivity, the evaluators have not been shown the identities of the MT systems throughout the experiment. All the systems have been shown as code names only, both when the evaluators have scored the output and analyzed their evaluation results, as seen in Table (4.5).

**Table (4.3). The Number of Words, Expressions and Terms Inserted by the Researcher in each Dictionary of the MT Systems.**

<table>
<thead>
<tr>
<th>Text Type</th>
<th>AL-Mutarjim AL-Arabey</th>
<th>An Nakel AL-Arabi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological</td>
<td>74</td>
<td>21</td>
</tr>
<tr>
<td>Chemical</td>
<td>57</td>
<td>32</td>
</tr>
<tr>
<td>Computational</td>
<td>50</td>
<td>7</td>
</tr>
<tr>
<td>Financial</td>
<td>78</td>
<td>26</td>
</tr>
<tr>
<td>Legal</td>
<td>195</td>
<td>9</td>
</tr>
<tr>
<td>Medical</td>
<td>93</td>
<td>31</td>
</tr>
<tr>
<td>Military</td>
<td>150</td>
<td>34</td>
</tr>
<tr>
<td>Petroleum</td>
<td>47</td>
<td>27</td>
</tr>
<tr>
<td>Physical</td>
<td>75</td>
<td>17</td>
</tr>
<tr>
<td>Social</td>
<td>85</td>
<td>18</td>
</tr>
<tr>
<td>Commercial</td>
<td>100</td>
<td>25</td>
</tr>
<tr>
<td>Water Engineering</td>
<td>77</td>
<td>26</td>
</tr>
</tbody>
</table>
4.2.3 Evaluators

In general, ratings typically are not highly reliable, and although methods are available for increasing the reliability a good dose of measurement error usually remain. A cardinal way to increase reliability of results is to employ multiple raters and average their responses.

Assessment of MT should be carried out by experts who:
1. know both SL and TL,
2. have specialized knowledge enabling them to judge the technical accuracy of the translation and
3. are sufficiently well acquainted with MT to evaluate the translation taking due account of the system’s potential and limitations (Van Slype, 1979b:156).

Table (4.4). Sources and Costs of MT Systems under Evaluation.

<table>
<thead>
<tr>
<th>Product</th>
<th>Market</th>
<th>Version of MT System</th>
<th>Local Price</th>
<th>Global Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL-Mutarjim AL-Arabey</td>
<td>AL-Misbar Computer Bureau/local</td>
<td>3.00/2002 Original</td>
<td>$100</td>
<td>$250</td>
</tr>
<tr>
<td>Golden AL-Wafi</td>
<td>AL-Misbar Computer Bureau/local</td>
<td>1.00/2002 Original</td>
<td>$35</td>
<td>$75</td>
</tr>
<tr>
<td>An Nakel AL-Arabi</td>
<td>Any local Computer Bureau</td>
<td>2.00/2000 Copy</td>
<td>$1</td>
<td>$600</td>
</tr>
</tbody>
</table>

Table (4.5). Identification of the MT Systems.

<table>
<thead>
<tr>
<th>Code</th>
<th>MT System Name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
On the basis of these elements, the researcher and a Ph.D. holder in MT at the Department of Translation/College of Arts/AL-Mustansiriyya University, act as evaluators in the present work. They are both fully aware of all the requirements of this evaluation task, i.e., how to apply the methods, metrics and scoring scheme.

To be less biased and to obtain more reliable results, each evaluator has worked separately after being given coded outputs of the MT systems evaluated. In other words, each evaluator has received the same set of a ‘side-by-side’ print-out of each text type which displays the ST and its three coded MT outputs in addition to a reference HT (assumed as a perfect translation) for comparison purposes. The ‘side-by-side’ print-out also provides information on how the ST has been analyzed and what expressions and domain-related terminologies have been highlighted for terminology assessment, which is so essential here.

It is worth stressing also the fact that the evaluators have made summative judgements not only over attributes of functionality, but also over some computational characteristics of the three systems. The results of the two evaluators are finally subjected to statistical analysis (See the next chapter).

### 4.2.4 User and Task Description

“Users” may be interpreted as most directly meaning the users of interactive software. They may include operators and indirect users who are under the
influence of or dependent on the use of the software. The actual operator of software will be called the ‘end-user’ and the person or organization to which the evaluation is addressed will be called the ‘customer’ of the evaluation (EAGLES, 1996:11, 23 and 101).

User-centered evaluation is clearly adequacy evaluation, which involves finding out whether a product satisfies the user needs. But, users are very numerous, and have widely differing needs. However, it should be possible to identify classes of users and to construct profiles of each one of these classes. These profiles can then be used as the basis for determining what attributes of particular classes of products are of interest to particular classes of users. Then, for each attribute, a procedure can be specified for discovering its value in the case of any particular product.

According to the ISLE Project (FEMTI, 2003:13-19), the characteristics of users are covered in three senses:

1. MT user, i.e., the end user who will interact with the MT system.
2. Translation consumer, i.e., the end user of the final product of the translation process.
3. Organizational user, i.e., the organization deploying the MT system.

In the present work, the end-user is an MT system user who produces the translation and directly interacts with the MT system or with the raw output produced by the MT system. This user may be a translator, a post-editor or a non-specialist user. In this case, formal linguistic education, proficiency in SL and TL and ease in computer use and manipulation are required. The user of the MT systems under evaluation is interested in producing high quality MT output representing different scientific and humanity domains. The chosen MT systems have been evaluated with the evaluation requirements (See 3.3.1) and the user’s tasks in mind. The high quality
output is expected to represent accurate, fluent, readable and understandable translation. Not only this, but the quality of output is also determined by the consistent and correct use of terminology and correctness on the syntactic and morphological levels.

The user in the present evaluation experiment aims at testing the adequacy and performance of each of the three MT systems with respect to the following tasks:

1. To what extent is the output of an MT system accurate, fluent and readable?
2. How precisely does the system translate domain-specific terminology?
3. Does the MT system produce a syntactically and morphologically correct output? How well-formed is the output?
4. In which domain does the MT system give the best output?
5. Which domain gives the worst output?
6. Does the MT system handle all the text types chosen acceptably well? and which system among the others is the best in this respect?
7. How fast is the system in performing the translation of each text?
8. How easily can the system be adopted to different specified environments?
9. What is the size in Kb of the text or document that the MT system can translate and display in its English and Arabic text boxes?
10. How easily can the system accept the addition of new words, dictionary building and other similar tasks?

The analysis of the above-mentioned user’s tasks has a serious impact on the characteristics the researcher has chosen to measure as part of the evaluation described in this study. For example, the researcher has decided
that faithfulness of translation (fidelity) and the clarity of meaning are of major importance, since the user must rely on the MT system’s output for the task of accurate transmission of scientific or legal content. For very similar reasons, the correct and precise coverage of terminology is of great importance. The researcher also believes that the syntactically and morphologically well-formed output has a positive effect on the comprehensibility and accuracy of information included in the text types evaluated. Other ISLE measures for MT systems in this study, i.e., the non-functional attributes like ‘speed of translation’, ‘storage’, ‘adaptability’ and ‘flexibility’ of MT system; are of less importance than the functional ones.

Thus, based on the user and tasks described in this section, the following five ISLE characteristics are identified as the most relevant to evaluate: 1.) readability, 2.) fidelity, 3.) terminology, 4.) syntax and 5.) morphology.

4.2.5 A User-Oriented Model of Test Types

The user and task-based requirements analysis described in the previous section leads to the identification of a set of attributes whose value is potentially of interest in the context of evaluating the products under consideration in order to assess their potential utility to a user or a class of users.

This section aims at providing a comprehensive and theoretically sound definition of types of test which allow the judgement of the performance of a piece of software as seen from the eyes of the user. Accordingly, there are three major test types: a.) scenario tests, b.) systematic tests and c.) feature inspection (EAGLES, 1996:73). Relevant to the present evaluation are ‘systematic tests’ merely.
4.2.5.1 Systematic Testing

Testing is considered as an important area in that quality has to be evaluated and consolidated. Under the term ‘systematic testing’, all testing activities will be subsumed that examine the behaviour of software under specific conditions with particular results expected. Systematic tests can be performed solely by software engineers and/or user representatives. There are three objectives that are particularly relevant for this type of testing (ibid: 32 and 80):

i. Examining whether the software fulfils pre-defined tasks.
ii. Examining whether the functions offered work properly.
iii. Examining the performance of the system.

Accordingly, user-oriented systematic testing is split up into: a.) task-oriented testing, b.) menu-oriented testing and c.) benchmark testing (ibid: 81). Relevant to this experimental work are the first and the last types only.

4.2.5.1.1 Task-Based MT Evaluation

Within the framework of adequacy evaluation, ‘task-oriented testing’ is performed to examine whether and to what extent a particular MT system meets some pre-defined tasks (See 4.2.4). The primary quality characteristic under investigation in task-oriented testing is functionality. The testing environment is normally the working place of the evaluator and is in principle not relevant to the interpretation of results. This type of testing can be carried out during the software development process at any stage of the software life cycle as well as with any ‘off-the-shelf’ software product. Apart from the technical environment of the evaluator (i.e., hardware and
software), no extra investment in testing equipment or instruments is normally necessary (ibid: 34, 81 and 82).

Since the major purpose in task-oriented testing is to assess the overall functionality of the MT systems by means of relevant data inputs, as well as to examine the quality of the data outputs, i.e., it evolves from the tradition of ‘black-box’ evaluation (Vanni and Miller, 2001), it is this type of tests that the researcher has found most relevant to the present evaluation. The goal is to determine what a system ‘gets right’ in its output such that a human translator can perform a specific task with it.

In order to evaluate the MT systems’ functionality, the task-based evaluation adopted has employed standard, rather than randomly-chosen, tests of MT output quality selected from the FEMTI (2003) with respect to the methods, metrics and scoring schemes. This is to estimate the values of the functional criteria of each MT system considered in terms of a black-box comparative and adequacy-oriented evaluation.

Finally, the metrics applied in task-oriented testing deliver mostly boolean (i.e., presence or absence of functions), quantitative (i.e., number of steps) or classificatory (i.e., how well a function performs a task) values. The last type of metrics is adopted here, which mostly fits the methods followed in each evaluation task in this work. Further, task-oriented testing is normally performed by only one (better two) evaluators and thus has to be seen in the light of subjectivity (EAGLES, 1996:82).

4.2.5.1.2 Benchmark Testing

Benchmark testing examines the performance of systems. The notion of performance can be applied to individual functions, to system modules or to the system as a whole. In the strict technical sense, a benchmark test is a
measurement of system performance which cannot be affected by variables resulting from human involvement. Typical reporting instruments applied for benchmark tests are checklists. When performing the same benchmark with different systems, it has to be kept in mind that both system parameters and environment variables are kept constant. The comparison of the benchmark results only makes sense if different MT systems, for instance, have access to the very same background material and are tested with the very same test text (ibid.: 34 and 84).

The major quality characteristic assessed by means of benchmark tests is ‘efficiency’ (ibid.). In the present study, some non-functional criteria, (i.e., speed of translation and storage) are evaluated in terms of special methods of system’s performance. Others, like ‘adaptability’ and ‘flexibility’ are tested by means of benchmark tests using a special checklist designed for this purpose.

It is worth noting that all of the types of systematic testing described here are potentially of interest in the context of adequacy evaluation. Task-oriented testing can indicate whether a product does, in fact, do what a user wants it to do. As for benchmark testing, the aim is to check whether a product meets a user’s minimum requirements. In this evaluation experiment, the non-functional, i.e., computational criteria are tested by a comparative performance and adequacy-oriented evaluation. It is important to emphasize that the goal is to test the systems’ capabilities not the hardware used.

4.2.5.2 Testing Instruments

According to EAGLES (ibid: 35, 85 and 86), the term instrument denotes any form of experimental data collection that can be used in combination
with the different test types (e.g., checklist with benchmark tests). The choice of instruments depends on various factors such as time, money constraints, testing experience, testing environment, etc.

In the testing phase, there are two major kinds of testing instruments, i.e., those that ask for a manual collection of data and those that perform automatic data collection. The most prominent manual testing instruments are questionnaires, checklists, interviews, observations and think-aloud protocols (ibid: 86). Among these, checklist is the most typical instrument applied for benchmark test adopted in the evaluation of some computational criteria in this study.

4.2.5.2.1 Checklists

Checklists are frequently used for manual software tests, particularly for any kind of inspection. Generally, the aim of using checklists is “… to obtain a concise and coherent description of the system in terms of objects, attributes functions, relations between objects as well as between objects and functions and estimated usability.” (Vainio-Larsson, 1990:325)

According to EAGLES (1996:87), checklists can deliver numerical, boolean and classificatory values. The most frequently used rating technique in checklists is availability rating, for which only the boolean values of yes/no (or true/false) are used. A more comprehensive rating score, normally with 5-7 items, is used for performance rating. Thus, for more objective and reliable rating score, performance scales of three and five features have been considered in the present evaluation of ‘flexibility’ and ‘adaptability’ (Appendix C).
4.3 Evaluation Procedure

The last step in the Evaluation Process Model involves three stages, namely: a.) measurement, b.) rating and c.) assessment. For measurement, the selected metrics are applied to the software product. The result is values on the scales of the metrics. In the rating stage, the rating level is determined for a measured value. Assessment is the final step of the software evaluation process where a set of rated levels are summarized. The result is a statement of the quality of the software product. Finally, managerial decision will be made based on the managerial criteria. The result is a managerial decision on the acceptance or rejection, or on the release or no-release of the software product (EAGLES, 1996: 65 and EAGLES, 1999:4).

In the next section of this chapter, the application of the metrics and methods selected for the evaluation of each of the quality characteristics under investigation is described in detail. The rating level of each of these criteria is also determined by the two evaluators.

The results for the measured values and the combination of values for each quality characteristic, the analysis of the existing correlation between the various scores of the two evaluators and the overall assessment of the performance of an MT system in terms of the functional and non-functional criteria (to be, then, compared with the other systems’ results) will be thoroughly analyzed and discussed in the next chapter.

4.3.1 Implementation of Evaluation Methodology

Prior to the process of identifying, classifying and discussing the results of the tests, there must be a clear and consistent method of analysis and evaluation on which the investigation of the quality characteristics and the overall system’s performance will be based. In this respect, the metrics and
methods directly associated with each quality characteristic under investigation (See 3.3.2.3) have been applied. The measures have been carried out in accordance with the standard metrics and methods. Different weightings have been assigned by each evaluator to the individual measures, in order to calculate the overall score. Moreover, the researcher has used descriptive statistical techniques and scoring schemes whenever required in the detailed evaluation task. To a certain reliability of assessment of these aspects, a second opinion is required. Thus, a specialist in MT has been approached and she has closely analyzed and evaluated the sample texts, using the same metrics and methods. Her judgements have almost been identical with the researcher’s.

The goal of this evaluation is two-fold:

a.) To measure the capabilities of each MT system with respect to the various quality characteristics considered.

b.) To find out which MT system is the best in terms of performing the tasks suggested earlier.

It should be emphasized that when it is perceived that a test on one aspect of the output would interfere with an evaluator’s ability to objectively assess a subsequent feature being evaluated; ordering of the tests is rearranged to avoid such interactions. For example, scoring of ‘fidelity’ is normally done after scoring for ‘intelligibility’ (or readability). Following the FEMTI (2003), the order of criteria to be evaluated in this study, matches the order of tests to be performed in the evaluation sequence. The result is a top-down ordering of tests.

Tasking this into account, the following features have been examined more closely and are ranked here according to their importance.
4.3.1.1 Testing of Functional Criteria

On the basis of the tasks relevant to the end-user’s needs in this study, only five functional quality characteristics have been investigated. These include ‘readability’, ‘fidelity’, ‘terminology’, ‘syntax’ and ‘morphology’. In this respect, the evaluators have carried out task-based tests of three MT systems based on ‘black-box’ approach. In testing these criteria, the evaluators are more interested in what each system can currently manage (i.e., declarative evaluation) and the extent to which a particular MT system meets the pre-defined user’s tasks (i.e., adequacy evaluation). The results obtained from this evaluation will serve translators who need MT systems which are easy to access and use, compatible with their computerized environment and work processes. The good MT system must enable them to make the best possible use of their expertise and experience and to increase the quality of translations they can do. In the last stage of these tests, comparative evaluation experiments have been conducted to determine the best system.

In this work, the black-box evaluation has been chosen due to the fact that commercial MT systems can only be evaluated by this approach (Volk, 2001). Consequently, there has been no access to the inner workings of these systems. Even so, it is desirable to be able to draw from such an evaluation enough conclusions about the various system components. In such a setting, the evaluation may not be able to pinpoint the error source; however, it will give an indication as to what sub-system is malfunctioning.

Accordingly, the researcher has suggested five questions designed to assess what each MT system under testing can do, in order to fulfil the pre-specified requirements of its users. So, the following are the questions on which the whole analysis of the MT outputs in every test has been based
(See the next chapter), each with a clear discussion of the procedure followed in its application.

4.3.1.1.1 Readability Rating

To what extent is the output of an MT system understandable, fluent and readable?

The above question motivates inquiry into the ease with which the output text can be read. In other words, it is an investigation of the degree to which some discernible meaning is conveyed by a sentence in the TT without any reference to the ST.

Four metrics have been decided upon for purposes of measuring ‘readability’ of the output text. The two evaluators have assigned a ‘clarity’ score to the same set of (268) output sentences representing 12 different domains. The same scoring scheme and procedure of evaluation is repeated with each MT system under investigation (three such systems are involved). For each TT sentence a score of 3, 2, 1 or 0 has been given, depending on whether a sentence is readable or not (See 3.3.2.3.1). The rating stage follows the careful reading of each TT sentence detached from its context. The evaluators have based their judgements on whether or not certain features are found in a sentence that lead to distortion of its meaning such as: a.) wrongly translated words or expressions, b.) unhappy choice of words, c.) missing word(s), d.) unknown word(s), e.) incorrect syntax and f.) incorrect order of words. Since systems have been given coded names, evaluators could not determine which sentence is from which system. The ‘readability’ score for each text is the mean of the scores of all sentences for both raters. The percentage value of this criterion for each text type has been calculated and compared with that of the other text types translated by the
same MT system to see which text type can be read and understood better. The same results have been compared on the level of the systems to determine which system translates a particular text type better for this criterion. Also, the average value of ‘readability’ of all text types for a particular MT system has been calculated and compared with that of the other systems to see which MT system has the highest performance in this respect. It is worth noticing that the total percentage value of ‘readability’ is calculated in terms of the correlation coefficient of the scores of the two raters.

After all, the value of ‘readability’ for an MT system in terms of the relative importance of this criterion in the calculation of the overall GSQ has been found.

Results of the analysis, evaluation and comparison are then tabulated and the necessary statistical requirements are computed taking into account variation across the raters’ judgements.

Besides, the aim of this question is to investigate the reasons behind the wrongly and ill translations produced by the systems considered in this study. These will be identified and discussed in detail in the next chapter.

4.3.1.1.2 Fidelity Rating

To what extent is the content of the source language sentence preserved in the translation?

Fidelity is designated the most important characteristic for the translator’s task. It is a measure of the information successfully conveyed from the SL text to the TL output. As the evaluators have good knowledge in the SL (English), this test has been performed by judging the ‘fidelity’ of the translation with respect to the original text only. Note that the HT is needed
here as a reference-adequate translation to detect syntactic and semantic errors in the MT output that have badly affected the ‘fidelity’ of TL texts. It is also used to help raters making more precise and accurate judgements and scorings. Further, it provides correct translation of output to support the discussion of sources of problems with ‘fidelity’ in the next chapter.

Fidelity scores for each text in this evaluation experiment are computed in the following manner: Each sentence has been assigned a value from a 4-point scale. These individual sentence values have then been averaged over the whole text (See 3.3.2.3.2). This is essentially the test proposed in Van Slype (1979b). The scoring is performed with values ranging from 0 to 3, based on the amount of information in the MT output after being compared with the original input text. The numbers of texts correspond to these in the ‘readability’ test using the same MT systems.

It is worth mentioning that the two evaluators are blind to which system produces the output. None is expert in the subjects covered by the sample texts (this is also the case with the other tests of the functional criteria). They have examined only the raw MT outputs of the systems considered consulting as reference the HT and rated how well the content of the original text is preserved. Losses of information, interference and combination of these factors have been taken into consideration in the evaluation process. For each text in this test, the total score of ‘fidelity’ is calculated with respect to the correlation coefficient of the scores of the two raters.

Like in the ‘readability’ testing, the highest rate of ‘fidelity’ with respect to text type and MT system has been calculated. Additionally, the value of ‘fidelity’ for an MT system in terms of the relative importance of this criterion in the calculation of the overall GSQ has been found. Details of the results are discussed in the next chapter.
4.3.1.1.3 Domain Terminology Rating

How precisely does the system translate subject-matter terminologies?

The correct translation of the domain-specific terms is one of the most important tasks an MT system is expected to perform in this test. Untranslated terms, wrongly translated terms, synonym usages, the occurrence of forms irrelevant to the TL and other issues have been taken into consideration in the evaluation of this criterion. A human translated text for each SL text acts as a guideline for raters, so that they consider anything other than an exact match as wrong even though the variations encountered may be correct from the language perspective.

The metric is the percentage of correctly translated terms. For this test, the researcher has designed user dictionaries and preference dictionaries for “AL-Mutarjim AL-Arabey’ and only user dictionaries that require insertion of grammatical and semantic data for each entry for ‘An-Nakel AL-Arabi’ system. This helps in increasing the potentiality of these MT systems in handling the exact and precise translation of the domain-specific terms and expressions. Such a facility is not available in ‘Golden AL-Wafi’ MT system. Therefore, it depends in its translation on the general dictionary merely. Testing the effect of these dictionaries on the lexical capabilities of the MT systems investigated is one of the aims of the present evaluation.

In the implementation of the terminology test, each domain term in the input has been highlighted. The same thing has been done in the human translated text to determine whether such terms properly occur in the MT texts. Specific dictionaries and domain-specific texts have been used for this purpose. As a final step in the evaluation procedure, the evaluators have counted the percentage of correct cases. As mentioned in (See 3.3.2.3.3), the
evaluators have assigned one point for each term for which the MT output matches its HT. No point is given otherwise. Then, the ratio of the number of domain terms appearing correctly in the MT output of each text under evaluation is calculated for the two raters. A further step in this evaluation task has involved the computation of the ratio of the number of precisely translated domain terms with respect to each MT system. Besides, the value of ‘terminology’ for an MT system in terms of its relative importance in the calculation of the overall GSQ has been found. The results of this test are tabulated and discussed in detail in Chapter Five.

4.3.1.1.4 Syntax Rating

To what extent does the MT output respect the reference grammatical rules of the TL? How syntactically well-formed is the output?

Two scoring schemes and methods have been suggested in this study (See 3.3.2.3.4) to evaluate the syntactic correctness of the MT output. Since it is more widely used in MTE literature and consumes less time in doing the judgements, the researcher has suggested adopting the second procedure of evaluation. This is based on the human rating of MT sentences on a 5-point scale. It is worth mentioning that although HT texts have been translated and revised by experts in the subject matter of the sample text and in translation, an expert in Arabic (the TL in this study) is also consulted to do the necessary corrections, if any, and the final editing and revision processes\(^5\) in these texts.

After the MT output of each text has been divided into numbered sentences (each is lined up with its equivalent reference HT), each evaluator starts grading each sentence on a scale of one to five, where five represents a sentence is perfectly grammatical. The syntax score for each sentence has then been calculated and the percentage value of the syntactic score for each
text type and MT system is also found with respect to the correlation coefficient of the scores of the two raters. This is to find the best syntactic performance on these levels for the purpose of comparison among text types and MT systems. After all, the value of ‘syntax’ for an MT system in terms of its relative importance in the calculation of the overall GSQ has been found. A detailed analysis, discussion and tabulation of results are presented in the next chapter.

4.3.1.1.5 Morphology Rating

Does the MT system produce a morphologically correct output?

The aim of this test is to investigate the degree to which the MT output respects the inflectional morphological rules of the TL. The method suggested in (3.3.2.3.5) has been adopted to calculate the morphological score of each sample text and MT system in this evaluation. The inflected words have been identified and underlined in the reference HT of each text type. This facilitates the work of the raters in detecting the morphological errors and making the corrections needed in the MT output with respect to number, gender, case, tense, aspect, etc.

In the present evaluation experiment, the raters have counted and underlined the inflectable words in the MT output of each sample text. Then, they have started with making all the morphological corrections in the output texts. The morphological score of each sample text has then been calculated as the ratio of the number of morphological corrections to the inflectable words in the MT output. Further, the final morphological score is calculated as the average value of all the texts under evaluation for each MT system with respect to the correlation coefficient of the scores of the two raters.
Finally, the value of ‘morphology’ for an MT system in terms of its relative importance in the calculation of the overall GSQ has been calculated. Details of this evaluation experiment will be explained in Chapter Five.

4.3.1.2 Testing of Non-Functional Criteria

Four computational (non-functional) criteria of MT quality have been tested in this evaluation experiment. These are: ‘speed of translation’, ‘adaptability’, ‘storage’ and ‘flexibility’. Among the other non-functional quality characteristics, these have been selected for their close relevance to the translator’s tasks. Benchmark tests using a checklist designed especially for this purpose have been carried out by the researcher and a computer expert to:

1. determine whether any of the products under investigation in a given application domain is adequate for its potential tasks, and if so, whether any of them is obviously more suited than the others (adequacy evaluation and comparative evaluation) and

2. measure the system’s performance when fulfilling certain translation tasks, such as: a.) speed of translation with respect to different text types, b.) adaptability to different software and hardware working environments, c.) storage of large texts and documents and d.) the flexibility of a particular MT system with respect to the addition of new words and expressions and/or dictionaries.

It should be emphasized that testing the non-functional criteria has achieved the greatest possible degree of objectivity and the results of the two evaluators seem completely identical.
Similar to the procedure followed in testing the functional criteria, the researcher has formulated four questions upon which the evaluation of the performance of the three MT systems considered is based. The sub-sections below present the details of testing each of the computational criteria involved.

4.3.1.2.1 Testing the Input – to – Output Translation

How fast is the MT system in performing the translation of each text?

The ultimate purpose of the translation task determines the importance of the ‘speed’ factor. In general, the monitoring of a relatively large volume of texts and on-line translation require high speed. This is also important in the case of in-house dissemination when the translations are sent to other people sharing aspects of the culture, terminology and domain knowledge to some extent.

In this work, though the evaluation task aims at finding out how fast a particular MT system is and which system under testing is faster than the others in performing the translation of different special-domain texts, the speed factor is not very important in the calculation of the total performance (GSQ) of a system. As it has been mentioned earlier (See 3.3.2.4) more concentration is on the functional criteria in this study.

Since ‘AL-Mutarjim AL-Arabey’ and ‘Golden AL-Wafi’ can be registered on one computer only (Table 4.2), no information can be obtain with respect to the effect of some factors (i.e., hardware and software) on the speed of translation. In addition, translating a certain text by ‘AL-Mutarjim AL-Arabey’ requires the determination of the theme of translation, the user dictionary and the preference dictionary. Only the user dictionary is involved in the translation process with ‘An-Nakel AL-Arabi’ and no such
dictionaries can be designed for ‘Golden AL-Wafi’. Thus, the present evaluation experiment aims at finding out the effect of these dictionaries on the speed of the translation process.

Using a stop-watch, the two evaluators work separately in calculating the time which each text of the twelve selected domains takes in translation. Each measure has been repeated three times by each rater for constancy of results. No variation in such results has been noticed. After the translation time of each sample text has been expressed by the form (number of words/time of translation), the evaluators start computing the average value of speed of translation of each of the three MT systems tested. This results from calculating the average value of speed of the twelve texts translated by each MT system. Then, the percentage value of this criterion for each MT system has been found (one value only, since the results of the two raters are completely identical). This is obtained by dividing the average value of speed of each MT system by the maximum average speed scored by any one of the three MT systems (which is considered the ratio of the translation speed to a typical speed).

After all, the value of ‘speed of translation’ for a particular MT system in terms of its relative importance in the calculation of the overall GSQ has been found. The tabulation and discussion of the results is presented in the next chapter.

4.3.1.2.2 Testing Adaptability
How the MT system is adaptable to different specified environments? To what extent can an MT system be easily ported to different operational settings?
This question aims at investigating the operational setting of an MT system and the extent to which it is adaptable to different PCs, operating systems and other software it requires (if any). This inspection is based on the evaluators’ observations and tests during system operation. In this respect, benchmark testing using values to be assigned on a special checklist (Appendix C) has been adopted. Here, a rating score covering five features is used for performance rating.

In this test, only one computer has been used, since the two MT systems ‘AL-Mutarjim AL-Arabey’ and ‘Golden AL-Wafî’ can only be registered on one computer. These systems have been installed on the researcher’s PC by the agents of ATA Company and through a direct on-line contact with the manufacturers of the systems. As for An-Nakel AL-Arabi MT system, it has been tested on different PCs by the two evaluators.

As it is important to obtain percentage values of the results of this test for the estimation of the GSQ of each MT system considered and systems’ comparison, a special rating scheme has been used. A rating score covering five features (ranging between 0-4) has been designed for this purpose. Each rater in this experiment has given one value that matches a given feature in the checklist for the evaluation of the system’s adaptability (See 3.3.2.3.7). On the basis of this, the final percentage value of this criterion has been calculated for both raters with respect to each MT system under evaluation. This result is used to indicate the value of ‘adaptability’ in the calculation of the overall GSQ. All the results of this evaluation experiment will be tabulated and discussed in the next chapter.

4.3.1.2.3 Testing Storage
What is the largest size in Kb of the text or document the MT system can translate and display in its English and Arabic text boxes?

As it has been mentioned earlier in Chapter Two of this study, two MT systems AL-Mutarjim AL-Arabey and Golden AL-Wafi are uni-directional, i.e., they translate English texts to Arabic only and not the other way around. According to ATA Software Technology (2002:2) and ATA software (2003 a and b), these two systems translate English texts that should be not more than 40 Kb (about 20 pages and 5,000 words) which is considered very good, i.e., there is a 40 Kb limit on displaying text in the text boxes English and Arabic. An- Nakel AL-Arabi MT system, on the other hand, is a bi-directional, i.e., it translates English texts into Arabic and vice versa. The manufacturing company (Cimos) of the system has not given any indication of the size of text or document the system can display and store in its text boxes. So, the present test aims at investigating the system’s storage capacity and the size of texts in Kb they can translate. It is also an attempt to find out whether the limits of the length of the text given by the manufactures are identical with these arrived at by the evaluators.

To translate documents larger than 40 Kb, both AL-Mutarjim AL-Arabey and Golden AL-Wafi have the option of direct or batch translation in the background (See Figs. 4.2 and 4.3, respectively). Both the English (ST) and the Arabic (TT) can be saved in one file making one document or in separate files with the same name, but the TT takes the extension (.ARB) which is Arabic text in Windows format.

For this evaluation experiment, English texts (.TXT) files of different sizes (e.g., 29.3 Kb, 37.7 Kb, 57.5 Kb, 60.6 Kb and 72.1 Kb) have been prepared. They all represent general texts, i.e., collection of most of the special-domain texts in this study. These general texts have been inserted
one after the other as input data in each MT system under testing to find out the largest size of text or document the system can display in its text boxes. As for direct translation of AL-Mutarjim AL-Arabey and Golden AL-Wafi, very large general texts (e.g., 382 Kb, 1.45 Mb, 2.91 Mb, 5.84 Mb and 23.3 Mb) have been prepared and tested. All the results of these evaluation tasks are tabulated and discussed in the next chapter.

Further, for each system under consideration, the evaluators have investigated the extent to which the system meets the pre-defined requirements, as observed during the benchmark tests. To obtain the final percentage value of ‘storage’ for each MT system under evaluation, the largest size of text in Kb it can translate and display in its text boxes (which represents the storage value of the system) is divided by the typical storage suggested in this study (i.e., the highest storage in Kb that any of the three Arabic MT systems can achieve).

Finally, the value of this criterion for each MT system in terms of the relative importance of ‘storage’ in the calculation of the overall GSQ has been found. This helps in comparing the total systems’ scores to see which system best satisfies the user’s needs.

**4.3.1.2.4 Testing Flexibility**

Is there a possibility of adding new words and dictionaries to the MT system?

The present test aims at investigating the system’s extensibility, i.e., how easy it is for a user to add new words/expressions and build new dictionaries. This possibility has a direct influence on the performance of translation and the overall quality of the MT system.
In order to provide users with a working system adapted to their environment, many translation products provide add-on dictionaries in certain subject areas and languages. Linguistic resources may also include the ability to create other bilingual or multilingual dictionaries to provide terminology quickly.

Figure (4.2). Direct Translation Dialogue Box in AL-Mutarjim AL-Arabey 3.00.
Regarding the three MT systems under testing, there is a possibility of free Internet updates for the most recent versions of AL-Mutarjim AL-Arabey and Golden AL-Wafi. This mainly concerns the systems’ dictionaries, i.e., their general dictionary and all the specialized science dictionaries (See Figs. 4.4, 4.5, 4.6 and 4.7, respectively). Such updates often concern the addition of new words, expressions and special-domain terminologies.
Figure (4.4). Bi-lingual Dictionary of Golden AL-Wafi 1.00.

Figure (4.5). Specialized Dictionaries of Golden AL-Wafi 1.00.
Unlike Golden AL-Wafi, AL-Mutarjim AL-Arabey gives the user the chance of creating and adding new dictionaries to the system. These include:
a.) user dictionaries, b.) the abbreviations dictionary and c.) preference dictionaries (See Figs. 4.8 and 4.9, respectively). These dictionaries can be in use when the programme is running at any one time to enhance the quality of translation. For new words and special-domain terms, the user can open a special user dictionary to enter these terms and give them their Arabic equivalents. It is important to notice that a user can create as many user dictionaries as he likes. Each dictionary can hold up-to 300 terms of 60 characters in length for English and the same for Arabic. When adding new words or terms to this dictionary of AL-Mutarjim AL-Arabey, there is no need to indicate the grammatical category or any other linguistic information of these words or terms. This makes it easy for users who are not expert in linguistics and translation to build their own dictionaries. The abbreviation dictionary is for abbreviations (e.g., P.O. Box = Post Office Box) only. This dictionary can hold up-to 300 entries too. Only the abbreviation and its equivalent in English should be entered and AL-Mutarjim AL-Arabey will equal the abbreviation to the text and translate it accordingly. As for the preference dictionary, the user of AL-Mutarjim AL-Arabey can build as many of these dictionaries as he needs. The preferred Arabic meaning can be selected for an English entry during the translation process. For example, the word ‘development’ has four Arabic meanings listed in this dictionary of the system as follows: (تطور – نمو – توسيع – تنمية) with the preferred Arabic word ‘تطور’. The user can change this to any of the four Arabic nouns according to the subject matter of the text and use it always.
Figure (4.8). User and Abbreviations Dictionaries of AL-Mutarjim AL-Arabey 3.00.

Figure (4.9). Preference Dictionary of AL-Mutarjim AL-Arabey 3.00.

The third MT system in this test, i.e., An Nakel AL-Arabi, works with the sentence as the basic unit of translation in a rule-based processing system and knowledge database. The translation memory database contains idioms, locutions and translated units (Fig. 4.10). Because of its deep sentence analysis and semantic connections ‘An-Nakel AL-Arabi’ is able to learn new rules and knowledge to be used later in the translation of new sentences. These rules and knowledge are added to the TM of the system after finishing
with the post-editing process of a given text where different translations of certain sentences are suggested. When translating, the software starts by looking in the TM for an identical sentence to the one found in the ST. It uses alignment techniques or similar approximation such as fuzzy matching. The translation memory can suggest close correspondences. The software can carry out replacements or invariable parts (e.g., numbers or proper nouns) or variable parts with declensions. After post-edition, the result can be stored as a potential collection of initial translation memories. Here, the English SL sentences and their Arabic translations in the TL can be added, deleted or modified from TM database.

Figure (4.10). Translation Memory in An -Nakel AL-Arabi.

In addition to TM database, An-Nakel AL-Arabi uses four dictionaries, i.e., general words dictionary, specific words dictionary, user words
dictionary and idioms dictionary. The general dictionary contains words in common use with approximately 100,000 words and locutions. The specific dictionary contains words used by specialists in a selected subject area. These are available as separate purchase and the user can enter and code his own specialized dictionary. An-Nakel AL-Arabi also enables users to create their own specialized dictionaries. The user dictionary contains words (e.g., nouns, verbs, adverbs, articles, etc) added or inserted by the user (Fig. 4.11). These words make the entries of the dictionary and they are introduced with some grammatical information about them. For example, the word ‘house’ needs all the necessary information about it like (noun, singular and feminine) to be added to the user dictionary. Also, if the word has different characteristics, the user may enter each of them, as in the examples below:

(1) I buy a car.  اشتري سيارة.  buy → V0

(2) I buy him a car.  اشتري له سيارة.  buy → V00

Figure (4.11). User Dictionary of An-Nakel AL-Arabi.
In this dictionary, the user can enter expressions like ‘from time to time’ and others that do not begin with a verb. Thus, users who have modest or no linguistic background find this system rather complex to deal with.

Users of An-Nakel AL-Arabi are able to create their own idioms dictionary too (Fig. 4.12). During translation, this system searches first the idiom translation in the user-defined idioms dictionary which is composed of expressions beginning with a verb (e.g., be able to). An-Nakel AL-Arabi’s approach is fully different from word-by-word translation. It is designed to support multiple meaning for words and to select the right meaning according to the assembling with other words in the clause. When translating phrasal verbs, for instance, the information about tense must be present simple, as in:

e.g.

<table>
<thead>
<tr>
<th>English</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Give back</td>
<td>رد + أعاد</td>
</tr>
<tr>
<td>Give in</td>
<td>استلم + اتفاق لـ + أقر بـ + سلم</td>
</tr>
<tr>
<td>Give off</td>
<td>أطلق + نشر + اخرج</td>
</tr>
<tr>
<td>Give on</td>
<td>اطل على</td>
</tr>
<tr>
<td>Give on to</td>
<td>اشرف على + اطل على</td>
</tr>
<tr>
<td>Give over</td>
<td>سلم إلى + كف عن + ينس</td>
</tr>
<tr>
<td>Give up</td>
<td>تنزل عن + تخلى عن + اقفل عن + ينس</td>
</tr>
</tbody>
</table>

When entering expressions like ‘give room’ to the idiom dictionary, ‘An-Nakel AL-Arabi’ is able to recognize all the equivalent expressions, as in:

e.g.
(3) The child **gives room to** his brother.

(4) My sister **gave birth to** a beautiful baby.

(5) The manager’s decision **gave rise to** a big strike.

(6) The little child **gave quarter to** the abandoned cat.

Figure (4.12). Idioms Dictionary of An-Nakel AL-Arabi.

In this evaluation experiment, the researcher has designed and built user and preference dictionaries for ‘AL-Mutarjim AL-Arabey’ and ‘An-Nakel AL-Arabi’ (See Table 4.3) taking into account unknown words, wrong translations and wrong forms (i.e., the source word is translated as a verb although it must be a noun). Since post-editing processes have not been
included in this study, the TM in An-Nakel Al-Arabi system is not tested. Thus, the two raters have based their judgments and scores on the extent to which it is possible for the three MT systems to accept the addition of new words, expressions and dictionaries merely. The systems’ capacity in this respect has also been taken into consideration.

The evaluation of ‘flexibility’ of the MT systems in this study involves the use of a checklist where a rating score covering three features have been used for performance rating. Results of the evaluators’ observations and testing of the systems’ capacities have been assigned on the checklist where each of these systems has been scored a certain value that stands for a certain feature. Then, for each MT system, the final percentage value of ‘flexibility’ has been calculated for the two evaluators on the basis of the checklist values and features (See 3.3.2.3.9).

Finally, for systems’ comparison, the value of this feature in term of the relative importance of ‘flexibility’ in the calculation of the overall GSQ has been found. Results of this test are tabulated and discussed in detail in the next chapter.

### 4.4 Summary

The previous discussion has given a full account of the present MTE experiment, where the quality of the three English-into-Arabic MT systems has thoroughly been tested. Concentration has been on the design and execution stages of the evaluation process. It has also attempted to give a description of key concepts in evaluation and outline some of its general characteristics.
Moreover, a discussion has been made of the evaluation framework focusing on the text sampling, the systems tested, evaluators, user and task description. The sample represents a total of 268 English SL sentences taken from twelve various special-domain texts. They have been used as input testing data in the evaluation of selected functional and non-functional characteristics of the three MT systems considered. These systems have been tested under experimental conditions. For more reliable results and evaluators to be less biased in their judgments, these systems have been given code names. In addition, the end-user (i.e., the translator who interacts directly with the MT system and its raw translation) and his tasks have been identified and discussed in relation to the ISLE characteristics of the translation quality.

Further, the chapter has presented a comprehensive and theoretically sound definition of types of tests and evaluations which allow the judgment of the total performance of the three MT systems from the viewpoint of the user. Accordingly, user-oriented systematic tests (i.e., task-oriented tests for the functional criteria and benchmark tests and the testing instruments for examining the non-functional criteria) of the MT systems have been applied. The last step in the Evaluation Process Model, i.e., the evaluation procedure with its three phases: a.) measurement, b.) rating and c.) assessment has focused on examining the quality characteristics under investigation by the two evaluators. The results for the measured values and the combination of values for each quality characteristic and the overall assessment of the MT systems’ performance will be analyzed and discussed in detail in the next chapter.

**Notes to Chapter Four**
(1) For limits of space (as the data extends over more than 60 pages), the sample texts in the present study have been saved on a CD-ROM where:

ST = Source Text in English
HT = Human Translation (as adequate translation)
GW = MT Output of Golden AL-Wafi System
MA = MT Output of AL-Mutarjim AL-Arabey System
NA = MT Output of An-Nakel AL-Arabi System

(2) The choice of the sample texts with respect to the various domains, sources of text sampling and number of sentences for the present evaluation experiment, has been confirmed of being “fine” through personal e-mails with outstanding figures in the field of MTE, like Dr. Ahmed Guessoum/Computer Science Department/University of Sharjah (Guessoum, 2002 d and e); Prof. Eduard Hovy / USC Information Sciences Institute (Hovy, 2002b) and Prof. Andrei Popescu-Belis / ISSCO/TIM/ETI / University of Geneva (2002a).

(3) Full information regarding the MT systems under evaluation (e.g., cost) has been obtained through the Internet and direct e-mails with the software companies of the products under evaluation. In the last e-mail (2002) with the manager of CIMOS, the researcher has been informed that a new version 3.00 of ‘An-Nakel AL-Arabi’ has been finished and released which is more enhanced and advanced. The price of this new product is $990.

(4) It is worth mentioning, here, that the researcher has received the two MT systems ‘AL-Mutarjim AL-Arabey’ 3.00 and ‘Golden AL-Wafi’ 1.00 on the 5th of February 2003, just after having finished the
experimental work with the old version of the first system. This requires the repetition of the experimental work, of course.

(5) The employment of ‘revision’ and ‘editing’ processes is a significant aspect for the improvement of any written text. The first involves making changes to the written text in order to clarify meaning such as changing lexis or reordering of words, logical connection between propositions and paragraphs, textual cohesive ties, explanation and elaboration of ideas and so on. Editing processes, on the other hand, concern themselves with making changes to the written text in order to correct syntax, spelling and punctuation (Hannouna, 1999: 221).
CHAPTER FIVE
CLASSIFICATION AND DISCUSSION
OF EVALUATION RESULTS
The last step of the evaluation procedure involves quality assessment of the software product. The present chapter, therefore, concerns itself with the presentation, classification and discussion of the results obtained by testing the three Arabic MT systems in terms of various criteria types.

The detailed results of each test and the overall quality assessment of the participating systems are presented and classified first. Then, some observations and problematic issues relevant to the evaluation results are discussed together.

5.1 Analysis and Classification of Results

The results of the evaluation process can be considered under two headings: a.) the variation in scores as a function of differences in sentences, criteria and raters; and b.) the average scores of the various translations and systems’ performance.

Accordingly, detailed analyses and classifications of the results concerning the various criteria types, i.e., functional and non-functional, are presented with Excel tables, charts and graphs. These are followed by the overall comparison of the individual systems in terms of quality assessment at both system and text level.

5.1.1 Evaluation Results of Individual Software Criteria

5.1.1.1 Functional Evaluation Results

As it has been stressed in the MTE literature, the criteria of ‘suitability’, ‘accuracy’ and ‘well-formedness’ are the most interesting from the user’s point of view in determining the quality of an MT output. In this study, the MT systems under investigation have been tested in terms of the classical
cognitive criteria such as ‘readability’ and ‘fidelity’, the more objectively measurable criteria like the consistent and correct use of terminology and the linguistic criteria, i.e., morphology and syntax. Here, the tests are designed to show what a system can currently manage and whether or not it can meet the pre-defined user’s tasks. A detailed analysis based on the black-box approach and adequacy/declarative evaluation of twelve various text types for each of the three Arabic MT systems reveals the following results.

These results are classified and presented on the basis of: a.) variation in scores between raters, b.) comparison of systems for text types and c.) interrelation among functional criteria.

5.1.1.1.1 Variation in Scores between Raters
The results of the application of the evaluation methods in testing the criteria of ‘readability’, ‘fidelity’ and ‘syntax’ (See Chap. 4) take into consideration the number of sentences for each grade on the scoring scale, the total sentences’ score value, and the average sentences’ score value with respect to each rater and each of the tested MT systems. The evaluation results are reported below as in Table (5.1) which shows the distribution of the scores obtained from the investigation of (268) sentences for each of the three quality characteristics and MT systems.

It should be noted that the total sentences’ score values (resulting from the summation of the sentences’ ratings of each score on the scale) are basically used in the calculation of the percentage values of the above-mentioned criteria for both raters. These percentages and their average values are demonstrated in Appendices D, E and F, respectively.

An overview of the numerical results including the ranking of the MT systems according to these results is presented below (See Table 5.1):
(1) Regarding the ‘readability’ test which aims at investigating the understandability and fluency of an MT output, the procedure of analysis relies on the use of a 4-point scale to obtain the final evaluation of each MT system (See Chaps. 3 and 4). Examination of (268) sentences for each MT system reflects the following:

a.) As for score (0) in the scale (indicating that meaning of sentence is not apparent even after some reflection), 45 sentences representing 16.79% of the total number of sentences in the ‘readability’ test have been graded so by the first evaluator (eval.1, the researcher in this study) and 53 sentences out of 268 that stand for 19.77% have been scored (0) by the second evaluator (eval.2) for ‘Golden Al-Wafi’ (GW) system. Results of rating the sentences of ‘Al-Mutarjim Al-Arabey (MA) manifest that only 4 sentences out of 268 constituting 1.49% have been given this grade by eval. 1; while eval.2 has scored 8 sentences out of 268 with (0). The percentage of this result is 2.9 of the total number of sentences in this test. For An-Nakel Al-Arabi (NA), 20 sentences representing 7.46% of the total number of sentences have got score (0) by eval.1, and 24 sentences that stand for 8.9% of the total number of sentences have been graded (0) by eval.2. GW, thus, has the highest number of sentences for score (0) among the other MT systems in the present evaluation; whereas MA has the lowest, i.e., very few sentences in the output of this system have no clear meaning.

b.) For the three MT systems, the ‘readability’ results show that both raters have graded the majority of sentences in this test with scores (1) and (2). This reflects that: a.) either some meaning can be gleaned from the sentence with some effort (i.e., less than 50% of the meaning is clear) or b.) meaning of the sentence is clear only after some reflection (i.e., more than 50% of the meaning is clear as a result of inferring it from the context or
reading the sentence several times). Accordingly, 97 sentences representing 36.19% of the total number of sentences in this test have been given score (1) by eval.1; while 91 sentences out of 268 that stand for 33.95% have been graded so by eval.2 for GW. Only 85 sentences constituting 31.71% of the total number of sentences have got score (2) by eval.1, and 80 sentences out of 268 constituting 29.85% have been given this grade by eval.2 for the same MT system.

Concerning MA system, the results indicate that 60 sentences out of 268 representing 22.38% and 59 sentences that stand for 22.01% of the total number of sentences in the ‘readability’ test have been given score (1) by evals.1 and 2, respectively. Further, for the same MT system, the two raters have graded 114 sentences out of 268 constituting 42.53% and 110 of the sample sentences in this test that stand for 41.04% with score (2).

Results of NA point out to assigning 124 sentences that represent 46.26% of the total number of sentences score (1) by eval.1; while 115 of the sample sentences constituting 42.91% have been given the same score by

Table (5.1). Two Evaluators’ Scoring Results of Readability, Fidelity and Syntax.

<table>
<thead>
<tr>
<th>System Type</th>
<th>Evaluator</th>
<th>Readability Scoring Scale</th>
<th>Total Score Value</th>
<th>Average Score Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Golden Al-Wafi</td>
<td>1</td>
<td>45</td>
<td>97</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>53</td>
<td>91</td>
<td>80</td>
</tr>
<tr>
<td>Al-Mutarjim Al-Arabey</td>
<td>1</td>
<td>4</td>
<td>60</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8</td>
<td>59</td>
<td>110</td>
</tr>
</tbody>
</table>
The two raters have assigned 72 sentences representing 26.86% of the total number of sentences and 75 sentences out of 268 that constitute 27.98% grade (2).

On the basis of the above-mentioned results, it is obvious that among these MT systems, NA has the highest number of sentences for score (1); while MA has the lowest number of sentences for this score. However, for
grade (2), a completely opposite result can be seen, i.e., MA has the highest number of sentences; whereas NA has the lowest.

c.) With respect to the highest score in the scale, i.e., grade (3), results of GW reveal that 41 sentences out of 268 representing 15.29% have been rated with score (3) by eval.1; while eval.2 has assigned 44 sentences that stand for 16.41% of the total number of sample sentences the same score.

For MA system, 90 sentences out of 268 that constitute 33.58% and 91 sentences that stand for 33.95% of the total number of sample sentences reflect a perfectly clear meaning on first reading and are, therefore, given grade (3) by eval.1 and 2, respectively.

The two raters have evaluated 52 sentences that represent 19.40% of the total number of sentences in the ‘readability’ test and 54 sentences out of 268 constituting 20.14% with grade (3) for NA system.

It should be stressed, here, that although some sentences show a completely clear meaning on first reading, especially for GW and NA, the rate of sentences with no clear or partially clear meaning (i.e., some meaning can be gleaned from the sentence with some efforts) is higher for these systems than the sentences in the MT output of MA. For this last MT system, most sentences reflect either a perfectly clear meaning or clear meaning after some reflection (i.e., a relatively good readability).

d.) Table (5.1) clearly shows the individual sentences’ scores as well as the average sentences’ score values (i.e., the sum of grades given by one rater divided by the total number of sentences) obtained by computing the general MT systems’ performance to get understandable and fluent meaning for an end-user. Accordingly, the average sentences’ score values for both evaluators regarding GW, MA and NA are (1.5), (2) and (1.5) respectively.
This indicates that among the three tested MT systems, MA has achieved the best result.

The final evaluation of the ‘readability’ test manifests that for MA; in general, meaning of sentence is clear only after some reflection. Although this is a relatively good result, the meaning of most sentences is still not perfectly clear on first reading. As for the other systems, it is apparent (from the results in Table 5.1 and the percentages) that for most sentences in these systems, some meaning can be gleaned from the sentence with some effort. However, for a considerable number of sentences in GW and NA, meaning of sentence is clear only after some reflection.

(2) A look at Table (5.1) indicates that the results for ‘readability’ correspond fairly accurately to those for ‘fidelity’, although the grades given for the latter are generally worse than these for the former quality criterion. It is worth noticing that any variation between the ‘readability’ rating and the ‘fidelity’ rating is due to additional distortion of information, which can arise from the reasons mentioned in (3.3.2.3.2). The results below are obtained from the ‘fidelity’ test in the present evaluation:

a.) For GW system, grade (0) in the scale of ‘fidelity’ has been given to 60 sentences representing 22.38% of the total number of sample sentences and 48 sentences out of 268 constituting 17.91% by evals.1 and 2. Results of MA reveal that 11 sentences that stand for 4.10% of the total number of sentences in this test have been graded (0) by eval.1; while only 4 of the sample sentences that represent 1.49% have been given this score by eval.2. The two evaluators have respectively rated 35 sentences constituting 14.17% of the total number of sample sentences and 28 sentences out of 268 that stand for 10.44% with score (0) for NA.
As it is the case in the ‘readability’ test, GW system has the highest number of sentences for this score; whereas MA has the lowest since very few instances in the data of this system reflect completely unfaithful translation of the ST.

b.) In general, the majority of sentences for the three MT systems have been assigned grades (1) and (2) by the two evaluators. This indicates that the MT output is either ‘barely unfaithful’, i.e., less than 50% of the original information passes in the translation or ‘fairly faithful’, i.e., more than 50% of the original information passes in the translation. Accordingly, the two raters have respectively assigned grade (1) to 111 sentences out of 268 that constitute 41.41% and 128 sentences representing 47.76% of the total number of sentences in the ‘fidelity’ test for GW. As for grade (2), eval.1 has assigned 87 of the sample sentences that stand for 32.46% this score; while 80 sentences constituting 29.85% of the sample sentences have been given this grade by eval.2.

Regarding MA, 101 sentences that represent 37.68% of the total number of sentences have been graded (1) by eval.1; while 99 of the sample sentences that stand for 36.94% have been given the same grade by eval.2. The two raters have evaluated 138 sentences constituting 51.49% of the total number of sentences in this test and 145 sentences out of 268 that represent 54.10% with grade (2).

Results of NA system point out to assigning 139 sentences by eval.1 and 146 sentences out of 268 by eval.2 with grade (1). The percentages of these results are 51.86 and 54.47 of the total number of sentences, respectively. Further, the two evaluators have rated 86 sentences representing 32.08 and 90 sentences out of 268 standing for 33.58% with grade (2).
The above-mentioned results and percentages show that the majority of sentences in the output of GW and NA have got grade (1) for the ‘fidelity’ test (NA has the highest number of sentences); while MA has the lowest number of sentences in this respect. But, the reverse occurs for grade (2) as the majority of sentences for MA have been assigned this grade; whereas the other two systems have the lowest number of sentences here.

c.) In comparison with the evaluation results of ‘readability’ for the sentences obtaining grade (3), very few sentences have been given this score in the test of ‘fidelity’ (See 3.3.2.3.2 for possible reasons of this) in the case of all the participating MT systems. Concerning GW, eval.1 has graded 10 sentences out of 268 representing 3.73% with score (3); while eval.2 has assigned (12) sentences that stand for 4.47% of the total number of sentences in the ‘fidelity’ test a similar grade. With respect to MA, 18 sentences constituting 6.71% of the total number of sample sentences and 20 sentences out of 268 that stand for 7.46% have been scored (3) by the two raters, respectively. As for NA, 8 sentences that represent 2.98% of the total number of sample sentences have been given grade (3) by eval.1; while 4 sentences out of 268 constituting 1.49% have been similarly rated by eval.2. As mentioned above and confirmed by the results and percentages of the ‘fidelity’ test, very few sentences point out to completely or almost completely faithful translation of the ST for all the evaluated systems. Nevertheless, MA among the others has the highest result, which is still an indication of low fidelity.

d.) As manifested in Table (5.1), the average sentences’ score values for the two evaluators with respect to GW, MA and NA are (1), (1.5) and (1.25). These results are supposed to reflect the general performance of these
systems to faithfully preserve the content of the ST in the translated output for an end user.

It is noticeable that the average grades for the three MT systems range from (1) to (1.5) which, generally, point out to the passage of 50% or less than 50% of the original information in the translation. For one of the most important characteristics to the translator’s task, these values indicate low systems’ performance in this respect.

Further, despite the fact that the average sentences’ score value for MA is (1.5) which reflects that the translated output of this system ranges from ‘barely faithful’ to ‘fairly faithful’, the two evaluators have assigned the majority of sentences of this system grade (2), i.e., a fairly faithful translation where more than 50% of the original information passes in the translation. However, the majority of sentences for both GW and NA have taken score (1) (with the average values of 1 and 1.25, respectively) which means that less than 50% of the original information passes in the translation of these systems, i.e., poor fidelity.

The criterion of ‘syntax’ is the last quality characteristic where the procedure of using a scoring scale (here, of 5-points) has been adopted in its evaluation. Table (5.1) reports the results of the syntax test which aims at showing how syntactically well-formed is the MT output of the evaluated systems, as seen below:

a.) Concerning GW system, the analysis of the syntax test displays that 115 sentences out of 268 representing 42.91% and 116 sentences that stand for 43.28% of the total number of sentences have been given score (1) which refers to completely ungrammatical construction by evals.1 and 2, respectively. As for MA, eval.1 has rated 92 sentences constituting 34.32% of the total number of sentences in this test with score (1); while eval.2 has
assigned the same score to 78 sentences out of 268 that represent 29.10%. The two raters have graded 97 sentences that stand for 36.19% of the total number of sample sentences and 96 sentences out of 268 representing 35.82% with score (1) for NA.

The above-mentioned results indicate that for all the MT systems, a large number of sentences have been given the lowest syntax score. Among these, GW has the worst results; while MA has the best, although the system’s performance is poor in this respect.

a.) Regarding grade (2) in the syntax scale pointing out to almost ungrammatical constructions, i.e., suffering from serious problems; the results of eval.1 for both GW and MA are identical, as for both systems 67 sentences out of 268 constituting 25% have been scored so. With respect to eval.2, the results show only a slight difference between the number of sentences assigned this score for the two systems, that is, 64 and 73 sentences out of 268 representing 23.88% and 27.23% have been rated with score (2) for GW and MA, respectively. Results of NA system reveal that eval.1 has rated 97 sentences that stand for 36.19% of the total number of sentences with grade (2); while 104 sentences out of 268 constituting 38.80% have been graded so by eval.2.

From these results, it seems apparent that NA has the highest number of sentences for this score in comparison with both GW and MA whose results are almost identical in this respect.

c.) The results for grade (3) show that the two evaluators have given 65 and 68 sentences out of 268 constituting 24.25% and 25.37% this score for GW. As for MA, 87 sentences representing 32.46% of the total number of sentences in the syntax test have been rated with score (3) by eval.1; while 95 sentences out of 268 that stand for 35.44% have been given this score by
The two raters have almost graded the same number of sentences for NA with score (3). The results and percentages for evals.1 and 2 are 49 and 48 sentences out of 268 constituting 18.28% and 17.91.

On the basis of the number of sentences for grade (3), MA among the other systems has the best result; while MA has the worst. This is an indication of some reasonably grammatical fragments produced by these systems, but which still suffer from one or two major problems.

For both scores (4) and (5) in the syntax scale, very few sentences for the three participating MT systems have been graded so by the two evaluators. The results manifest that for GW, 16 sentences out of 268 representing 5.97% and 12 sentences constituting 4.47% of the total number of sentences have been rated with score (4) for evals.1 and 2. As for grade (5), the results are worse than these for grade (4) where only 5 and 8 sentences that stand for 1.86% and 2.98 of the total number of sentences have been given this score by evals.1 and 2.

Regarding MA system, 20 sentences out of 268 that represent 7.46 and 17 sentences constituting 6.34% of the total number of sentences reflect almost perfect grammatical structure which has one or two small problems and are, therefore, given grade (4) by evals.1 and 2. However, very low results and percentages are assigned by the two evaluators for grade (5). Accordingly, eval.1 has rated only 1 sentence that stands for 0.37% of the total number of sentences with this score; while eval.2 has given the same grade to 5 sentences out of 268 representing 1.86% merely.

The two evaluators have assigned 21 sentences that stand for 7.83% of the total number of sentences in the syntax test and 14 sentences out of 268 constituting 5.22% grade (4) for NA system. As for grade (5), 4 and 6
sentences out of 268 that represent 1.49% and 2.23% have been rated so by evals.1 and 2.

It should be noted, here, that the results of the three MT systems look almost identical for these two grades with the lowest number of sentences in comparison with the other syntax scores. The very slight differences between the evaluators’ rating for the systems involved are statistically insignificant.

It is noticeable from the results of the syntax test presented in Table (5.1) that the raters have given the evaluated systems the same relative ranking. In fact, the absolute scores between raters are very close. In looking at the average sentences’ score values, the results of the systems are identical, i.e., all of them have obtained an average value of (2) which points out that sentences are almost ungrammatical and suffer from serious problems. However, one of the major problems of syntax for these MT systems is that of word order (See 5.2.2.2). Despite the implications of this metric, in general, the result is worse than that is expected and is an indication of more than two major syntactic problems per sentence for the three systems, i.e., they perform poorly in this respect.

Nevertheless, regarding the ranking of systems for grades 1, 2 and 3 in the syntax scale, it is apparent that the majority of sentences for MA have obtained score (3) referring to reasonably grammatical sentences, but with one or two major problems. As for NA, the system has the highest number of sentences with grade (2) among the others reflecting almost ungrammatical constructions. Further, GW has the highest number of ungrammatical sentences, i.e., given grade (1), in comparison with the other two systems.
(3) It is worth stressing, here, that more objective methods have been adopted in the evaluation of the criteria of ‘terminology’ and ‘morphology’ (See 4.3.1.1.3 and 4.3.1.1.5) which lead to obtain results manifesting a complete correspondence between the two raters. Therefore, the percentages appearing in Figure (5.3) and Figure (5.5) in (5.1.1.1.2) refer to one value for each text type with respect to each MT system. On the basis of this, results of these two quality characteristics will be classified and presented in detail in terms of text types and general software performance (See the next sections).

However, having been evaluated using a scoring technique based on subjective scales, at least two raters should be involved for the reliability of the testing results. This concerns the criteria of ‘readability’, ‘fidelity’ and ‘syntax’. In order to find the level of correspondence in the results between the two raters, it is important to use the statistical widely used measure, i.e., the correlation coefficient value \( r \) (See Appendix G). This basically refers to measuring the relationship between two variables. In this study, the two variables represent the scoring values of evals.1 and 2 for these criteria. The obtained correlation coefficient values are based on Pearson’s linear correlation (Downie and Heath, 1983:103). The Excel results are tabulated in Table (5.2) below, where the input data for each value represents twelve values for each evaluator that stand for twelve text types (i.e., twenty-four values for each individual criterion and each MT system).

According to Downie and Heath (ibid.), in every usage an \( r \) of 0.8 and above is considered a high coefficient where all the points on the scatter gram tend to fall along a straight line. In addition, an \( r \) around 0.5 is considered moderate while that of 0.3 and below indicates a low coefficient.
Thus, the results of the tested systems in this study reflect very high correlation coefficient values (most of which are above 0.9) which means a statistical correlation. This points out to the fact that the scores of the two evaluators are in close agreement where a strong correlation exists; although for all the three tests not exactly the same sentences have been given the same grades by these raters. So, the results of only one evaluator can better be relied on in this respect than the average scores between the raters. In this study, the researcher’s (i.e., eval.1) results have been taken into consideration merely.

(3) It is worth noticing that the above-mentioned results could mainly be due to various general and specific factors which will be discussed in detail later (See 5.2).

### 5.1.1.1.2 Comparison of Systems for Text Types

For each of the functional criteria in this study, the performance of the individual MT systems has been investigated in terms of the various text types. This is to show in which domain a particular system is able to produce

<table>
<thead>
<tr>
<th>Criterion</th>
<th>MT System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Golden Al-Wafi</td>
</tr>
<tr>
<td>Readability</td>
<td>0.98</td>
</tr>
<tr>
<td>Fidelity</td>
<td>0.98</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.93</td>
</tr>
</tbody>
</table>
fluent, accurate and well-formed output for the end-user. The results of analyzing (36) sample texts to show the ranking of the MT systems involved with respect to these criteria and the relevant illustrative graphs are given below. It is worth noticing that the texts have been sequenced randomly, but in the same way in all these graphs\(^{(1)}\). Any change in their sequence will result in different graphic patterns without affecting the present evaluation. Thus:

a.) **Readability Test:** A look at Figure (5.1) and the assessment of ‘readability’ of the raw Arabic output of systems GW, MA and NA show that MA gives the best values for the majority of text types in comparison with the other MT systems’ results. This indicates a relatively good fluency and understandability of most of the texts evaluated. Accordingly, this system has obtained good results with the computational, legal, petroleum, commercial and water engineering texts (as the values of these are around 80%).

![Figure (5.1). Variation of the Three MT Systems’ Readability Based on Text Types.](image-url)
Generally speaking, the three MT systems have performed best with the computational and commercial texts (See Fig. 5.1). The percentages of the results for GW, MT and NA are (68.66, 81.66 and 76.66 for the computational text) and (80.12, 83.33 and 76.60 for the commercial text), respectively. But, on the other hand, these systems have obtained bad results with certain text types, as GW has got the worst value with the chemical text where only 30.55% of its sentences are clear and understandable. In addition, MA has obtained the lowest value with the same text type for only 48.44% of the sentences of his text reflect clear meaning. As for NA, the system has performed worst with the military text getting a value of 27.21% merely.

Further, the average value of the texts’ readability for each of the participating MT systems has been calculated (Fig. 5.1). The results reveal that MA has obtained the highest value of 69.48%; while both GW and NA have got the values of 48.54% and 52.77% respectively, where the texts of GW point out to the lowest readability in comparison with the other systems. In brief, among the other MT systems, the results of MA (with the exclusion of the chemical text) manifest a relatively good fluency and understandability for all the text types. So, the system can be suggested as a good tool for initial raw translation for the end-user who only seeks for clarity of meaning to understand the text on first reading especially with business and scientific domains.

b.) Fidelity Test: In general, the results of the ‘fidelity’ test are not encouraging, in fact. Although the analysis of the raw MT output for the three evaluated systems reveals that MA has obtained the best values for the majority of the text types, this result is still an indication of fair faithful translation of the content of the ST. Accordingly, in Figure (5.2), it seems
clear that the values of MA are the highest among the other participating systems with the exception of the chemical text as the value of this system is lower than this of NA where only 37.72% in comparison with 41.66% of the original information passes in the translation of this text. Moreover, the results of this system with both the social and commercial texts are lower than these of GW, although the differences are not big. The percentages of these results are (58.33 in comparison with 60) and (57.96 in comparison with 60.31).

Regarding the best performance of the three MT systems, Figure (5.2) shows that GW has got the highest values with the social and commercial texts. The percentages of these results are 60 and 60.31, respectively. As for MA, the highest values concern the medical, petroleum and legal texts, the percentages of which are 63.33, 61.28 and 60.33. With respect to NA, the best values are given to the legal and commercial texts as 53.96% of the sentences in the MT output of the former and 53.33% of these sentences in the latter are faithful to the original ST.

![Figure (5.2). Variation of the Three MT Systems’ Fidelity Based on Text Types.](image-url)
However, these MT systems have also obtained bad results with other text types. The result of the chemical text for GW point out to the lowest value of 15.72% among the other text types translated by this system. Moreover, MA has performed worst with both the biological and chemical texts. The percentages of these results are 36.36 and 37.72, respectively. The output of the military text for NA has the lowest fidelity among the other domains’ output, with the value of 20.24% merely.

On the basis of the average values of the texts’ fidelity, the percentages of GW, MA and NA are 39.22, 53.61 and 41.62, where MA has obtained the highest value and GW the lowest (See Fig. 5.2). These values are, in general, an indication of poor to fair syntactic and semantic fidelity of these systems that result in producing barely faithful TTs to the original STs.

c.) **Terminology Test:** First and foremost, it should be stressed here that the results obtained from the evaluation of domain-specific terms in the output of the tested MT systems are restricted to exact match with the HT for the selected terms in the SL texts. Consequently, for the reasons mentioned in (4.3.1.1.3), an overview of the results gives a bad indication even for the systems where the user can enter and code his own specialized dictionaries.
In general, for the three MT systems, Figure (5.3) clearly shows a big gap between the highest and the lowest values given to certain text types making the range between these values so great. Concerning the highest values obtained by each of these systems for exactly accurate translation of domain-specific terms, MA has performed best with the computational and military texts, where the percentages of these results are 60.97 and 67.27, respectively. As for NA, the system has obtained the highest results among the others with both the chemical and social texts. The percentages of these results are 68 for the former and 71 for the latter. With respect to GW, the highest result (also its only good result) obtained, here, is with the commercial text getting a value of 66.66\% in comparison with the other systems.

As it is the case in testing the other criteria, the participating systems have also obtained bad results with certain text types. Here, the lowest values for the three systems have been given to the petroleum text. The percentages of these results for GW, MT and NA are 28.12, 18.75 and 3.12, respectively. Although these values are bad for all the systems, GW has got the lowest one. The same system has also obtained a similar very low value with the water engineering text. The percentage of this result is 3.70 merely. With reference to Figure (5.3), the average values obtained from the ‘terminology’ test point to almost identical results for both MA and NA with only a very slight difference in the percentages of the two systems (i.e., 46.87 for the former and 46.17 for the latter). This ranks MA at the top with the highest value in comparison with the other systems. However, this result
indicates that the system fails to produce exactly accurate translation of domain-specific terms as the result in most cases is less than 50%. Further, very poor results have been obtained from the evaluation of GW, as the average value of most of the texts of this system is only 22.68 reflecting an apparent weakness in terminology translation.

d.) Syntax Test: In general, a thorough examination of the evaluation results of the syntax test manifest very close values for most of the text types with respect to the three MT systems. In most cases, the range is from 30% to 55%. In other words, no significant differences are noticeable in the values of these systems (See Fig. 5.4). Nevertheless, each of these systems gives the best and worst results for certain text types. Accordingly, MA system has shown the best results with the computational text with a percentage of 55, which is the highest among the other systems. Moreover, GW has also performed best with the same text type, but the value it has obtained is lower than that of MA, i.e., 53%. As for NA, the best value is given to the legal text with a percentage of 48. However, a look at these results gives the indication of poor grammatical correctness in the output of these MT systems for even the highest values they have obtained are around 50% or less.
As for the bad results of these systems, MA has obtained the lowest value with the biological text for a percentage of 34.54; which is very close to the results given to the other systems for the same text type, i.e., GW has got a value of 33.63% ; while that of NA is 32.72%. But, these two last systems have performed worst with other text types, as the lowest value obtained by GW is with the chemical text, while NA has the lowest value with the military text. The percentages of these results are 30 and 20.36, respectively.

All in all, the average values of the participating systems are very close, as the percentages of the results of GW, MA and NA are 39.75, 42.75 and 40.40 respectively. This means that the three systems have performed almost the same for all text types in the syntax test. Further, the low results of these systems explain the frequently bad output quality.

**Figure (5.4). Variation of the Three MT Systems’ Syntax Based on Text Types.**

As for the bad results of these systems, MA has obtained the lowest value with the biological text for a percentage of 34.54; which is very close to the results given to the other systems for the same text type, i.e., GW has got a value of 33.63% ; while that of NA is 32.72%. But, these two last systems have performed worst with other text types, as the lowest value obtained by GW is with the chemical text, while NA has the lowest value with the military text. The percentages of these results are 30 and 20.36, respectively.

All in all, the average values of the participating systems are very close, as the percentages of the results of GW, MA and NA are 39.75, 42.75 and 40.40 respectively. This means that the three systems have performed almost the same for all text types in the syntax test. Further, the low results of these systems explain the frequently bad output quality.

**e.) Morphology Test:** With respect to testing the degree to which words in the MT output of the TL are correctly inflected, the results of the three systems for the majority of the text types are very poor (See Fig. 5.5). The
The problem of morphology is, in fact, one of the main reasons of distortion of meaning and form of words that affects the quality of the whole text.

It should be noticed that not more than two text types for both MA and NA have obtained a value above 50% for this criterion. The results show that the highest value given to MA is 63.19% with the computational text; while NA has performed best with the computational and the legal texts, the percentages of which are 51.45 and 50.37, respectively. As for GW, the highest value the system has obtained is with the computational text too, but it is the lowest in comparison with the values of the other systems in this respect. The percentage of this result is 48.54 as all the text types with GW have got values less than 50%.

![Figure 5.5. Variation of the Three MT Systems’ Morphology Based on Text Type.](image)

Although, MA among the other systems has the best results in the ‘morphology’ test, the average values of the three systems reflect a serious defect in their morphological components resulting in producing Arabic MT
output of poor quality. Thus, the percentages of the results for GW, MA and NA in this respect are 29.54, 40.59 and 35.75, respectively.

Further, the average values of all the functional criteria in this study are graphically illustrated in Figure (5.6) below, where it seems clear that the average values of MA are the highest among the other systems and these of GW are the lowest with respect to all these criteria. This indicates that MA has the best functional performance; while GW has the worst. As for NA, the values it has obtained rank the system between MA and GW.

![Figure (5.6). Average Values of Functional Criteria for Arabic MT Systems.](image)

**5.1.1.1.3 Interrelation among Functional Criteria**

The correlation coefficients between the five functional criteria for each of the participating MT systems obtained by using the cross correlation matrix (which is also based on twelve texts’ values for each individual criterion) are shown in Tables (5.3), (5.4) and (5.5). These values are usually tested for
their significance by subjecting them to t-test which represents a measure of a standard error of r and should be computed by the following formula:

\[ t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \]

as t is a function of \( t_1 - \frac{\alpha}{2} \) where, \( \alpha \) is the level of significance, the most frequently of which is 0.05, 0.01 or 0.001.

Table (5.3). Correlation Coefficient Matrix of the Functional Criteria of Golden Al-Wafi

<table>
<thead>
<tr>
<th></th>
<th>Readability</th>
<th>Fidelity</th>
<th>Syntax</th>
<th>Terminology</th>
<th>Morphology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability</td>
<td>1.0</td>
<td>0.72</td>
<td>0.72</td>
<td>0.81</td>
<td>0.67</td>
</tr>
<tr>
<td>Fidelity</td>
<td>0.72</td>
<td>1.0</td>
<td>0.61</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.72</td>
<td>0.61</td>
<td>1.0</td>
<td>0.53</td>
<td>0.8</td>
</tr>
<tr>
<td>Terminology</td>
<td>0.81</td>
<td>0.73</td>
<td>0.53</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>Morphology</td>
<td>0.67</td>
<td>0.66</td>
<td>0.8</td>
<td>0.6</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table (5.4). Correlation Coefficient Matrix of the Functional Criteria of Al-Mutarjim Al-Arabey

<table>
<thead>
<tr>
<th></th>
<th>Readability</th>
<th>Fidelity</th>
<th>Syntax</th>
<th>Terminology</th>
<th>Morphology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability</td>
<td>1.0</td>
<td>0.74</td>
<td>0.68</td>
<td>-0.085</td>
<td>0.61</td>
</tr>
<tr>
<td>Fidelity</td>
<td>0.74</td>
<td>1.0</td>
<td>0.42</td>
<td>-0.21</td>
<td>0.27</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.68</td>
<td>0.42</td>
<td>1.0</td>
<td>-0.20</td>
<td>0.61</td>
</tr>
<tr>
<td>Terminology</td>
<td>-0.085</td>
<td>-0.21</td>
<td>-0.20</td>
<td>1.0</td>
<td>0.35</td>
</tr>
<tr>
<td>Morphology</td>
<td>0.61</td>
<td>0.27</td>
<td>0.61</td>
<td>0.35</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Table (5.5). Correlation Coefficient Matrix of the Functional Criteria of An-Nakel Al-Arabi

<table>
<thead>
<tr>
<th></th>
<th>Readability</th>
<th>Fidelity</th>
<th>Syntax</th>
<th>Terminology</th>
<th>Morphology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Readability</td>
<td>1.0</td>
<td>0.8</td>
<td>0.63</td>
<td>-0.134</td>
<td>0.11</td>
</tr>
<tr>
<td>Fidelity</td>
<td>0.8</td>
<td>1.0</td>
<td>0.8</td>
<td>-0.06</td>
<td>0.33</td>
</tr>
<tr>
<td>Syntax</td>
<td>0.63</td>
<td>0.8</td>
<td>1.0</td>
<td>-0.1</td>
<td>0.34</td>
</tr>
<tr>
<td>Terminology</td>
<td>-0.134</td>
<td>-0.06</td>
<td>-0.1</td>
<td>1.0</td>
<td>0.55</td>
</tr>
<tr>
<td>Morphology</td>
<td>0.11</td>
<td>0.33</td>
<td>0.34</td>
<td>0.55</td>
<td>1.0</td>
</tr>
</tbody>
</table>

In this study, the critical $r_c$ value (0.576) is directly taken from a developed table of values (Downie and Heath, 1983:329) depending on 5 percent level of significance and a degree of freedom $df = n – 2$ or 10 in the present work, since $n$ means the number of pairs used in computing $r$. Accordingly, it is unnecessary to go through any calculations.

A look at the correlation coefficient values for each system in the tables above points out that the $r$ values are compared with $r_c$ and it is statistically emphasized that each should have a value larger than (0.576) to be significant. In this respect, it is noticeable that for GW 88% of the $r$ values are significant indicating that the data of the interrelation among the functional criteria fit with Pearson’s model (ibid: 103) and are of linear relation. On the other hand, only 33% of the $r$ values are significant for MA; while 22% of these values are significant for NA. It is worth stressing, here, that there are often situations where the relationship between two sets of variables is not linear, and then the calculated $r$ is an underestimate of the true relationship between the two variables.
However, from the linguistic viewpoint, the five evaluated functional criteria are, in fact, highly interrelated. In other words, any deterioration in one affects the quality of the others. For instance, any syntactic, morphological or lexical defect in a sentence may lead to a total or partial distortion of its form and meaning, and, in turn, result in low fidelity and readability. Although such relations are very close, the resulting $r$ approaches zero.

Further, as an example, a deviation from linearity is apparent on a scattergram, made for the two variables (readability and terminology) of MA system (See Fig. 5.7). Here, it is found that the line is not apparent and the points are spreading evenly in all directions giving an indication of no relationship. Nevertheless, it is possible that a very high, but non-linear relationship will appear to be very low on the basis of Pearson’s model (ibid.).

Thus, according to the findings of this study, the bivariate relationship is considered curvilinear and the eta coefficient (i.e., correlation ratio) can be used, which reflects the variance accounted for by the best-fitting line, be it curved or straight. This is out of the scope of the present work and further research on this particular aspect may be desirable.

![Figure (5.7). Scatter Plot of the Readability and Terminology Results of MA System.](image)
5.1.1.2 Ranking of MT systems for Non-Functional Criteria

As mentioned earlier in this study (See 3.3.2.4), the top level tasks the participating systems are required to address are the functional (linguistic and cognitive) requirements. In addition, four non-functional (computational) quality characteristics have been investigated for the adequacy of system’s performance in fulfilling certain tasks relevant to the translation process (See 4.3.1.2). Therefore, benchmark tests have been carried out to show the central properties for product choice and to check the possibilities the individual systems offer for the end-user (adequacy/performance evaluation). Besides, the tests intend to demonstrate the differences among these systems and the systems’ ranking with respect to the stated criteria (comparative evaluation).

It is worth noticing that the results of the two evaluators (i.e., the researcher and a computational expert) are remarkably identical. These are tabulated and presented below.

5.1.1.2.1 Speed of Translation Results

In order to find out how fast a particular MT system is and which system under investigation is faster than the others in doing the translation of various text types, the speed of translation for each special-domain text and MT system is measured. Moreover, the average value of speed for each system is found and compared with these of the others. The results of such performance and comparative evaluation are presented below:

(i) In general, there is no significant relation between the resulted speed of translation and the system’s storage. However, it is noticeable that for long
texts (i.e., of large size in Kb), the speed of translation is higher than that with the ones smaller in size. This is rather obvious with GW and MA and not very much applicable to NA as the results in Table (5.6) illustrate. Here, general texts of different sizes representing various domains each are tested. On the basis of this, it is concluded that the larger the text, the faster its translation would be. Nevertheless, this aspect, in fact, still needs further research and investigation.

Table (5.6). Speed of Translation with Respect to Text Size.

<table>
<thead>
<tr>
<th>Text Size in Kb</th>
<th>Speed W/Sec For GW</th>
<th>Speed W/Sec for MA</th>
<th>Speed W/Sec for NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>545</td>
<td>182</td>
<td>732136</td>
</tr>
<tr>
<td>8</td>
<td>674</td>
<td>337</td>
<td>112</td>
</tr>
<tr>
<td>20</td>
<td>1002</td>
<td>445</td>
<td>138</td>
</tr>
<tr>
<td>40</td>
<td>1036</td>
<td>491</td>
<td>-</td>
</tr>
</tbody>
</table>

(ii) The time of translation for each sample text with respect to the three systems is also calculated. Table (5.7) manifests the performance time of each individual text type and system, in addition to the average time values in seconds. It seems clear the GW has the best performance among the others with an average time value of 1.25 second for the translation of each text. Regarding MA, it has the second position in the rank of systems with an average time value of 1.58 which is not a big difference from the result of GW. Indeed, NA takes more time in the translation of the sample texts which is in most cases more than that of GW and MA. Consequently, it has the lowest position in the rank of these systems with an average time value of 4.9 seconds per text.

(iii) As the manufacturers of ATA company have indicated (ATA Software, 2003 a and b), the speed of translation depends on the speed of the
PC (i.e., the CPU) and it is calculated by counting the number of words in any text under translation and the time it takes in performing this task. They have not given a typical speed of translation for the systems.

Since the storage of the participating systems is limited (i.e., 29 Kb for NA and 40 Kb for both GW and MA), the speed of translation of large texts with a storage exceeding 40 Kb cannot be measured (See 5.1.1.2.3 for more details). Thus, in the present evaluation, the typical speed of translation is found. This represents the maximum speed of 100% with the highest number of words translated per second (i.e., 855 W/Sec) for the petroleum text regarding both GW and MA (See Table 5.5). In the light of this result, the average values for the evaluated systems considering twelve text types for each, are: 526 W/Sec (61% for GW), 472 W/Sec (57% for MA) and 133 W/Sec (15.5% for NA). Obviously, results of GW and MA are almost identical except with the legal and commercial texts where GW has a higher performance (See Table 5.8). This table shows NA in the lowest rank among the other systems due to certain factors which will be discussed later.

Table (5.7). Translation Time for the Selected Texts.

<table>
<thead>
<tr>
<th>Text Type</th>
<th>Number of Words per Texts</th>
<th>Time of Translation in Seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Golden Al-Wafi</td>
</tr>
<tr>
<td>Biological</td>
<td>703</td>
<td>2</td>
</tr>
<tr>
<td>Chemical</td>
<td>570</td>
<td>1</td>
</tr>
<tr>
<td>Computational</td>
<td>458</td>
<td>1</td>
</tr>
<tr>
<td>Financial</td>
<td>484</td>
<td>1</td>
</tr>
<tr>
<td>Legal</td>
<td>1071</td>
<td>2</td>
</tr>
<tr>
<td>Medical</td>
<td>576</td>
<td>1</td>
</tr>
</tbody>
</table>
(iv) In order to investigate the effect of the various software and hardware requirements on the translation speed of the participating systems, a number of tests are carried out where these systems run on different PCs with various specifications. It is worth stressing, here, that such an experiment is only possible with MA (the old version 2.00) and NA, since both GW and MA (the new version 3.00) are highly protected and can be registered on one computer only.

The results of these tests are illustrated in Table (5.9) below. All the important features of the PCs such as processor speed, cache memory size, type of operating system, number and kind of software installed on the computers’ hard disk and the size of the hard disk are taken into consideration. It is found that the dominant parameter in determining the speed of translation of these systems is the processor speed as the table reflects.

Table (5.8). Speed of Translation and the Evaluation Speed % of the Selected Texts.
As there is a limit of displaying texts in the text boxes English and Arabic (See 5.1.1.2.3), the standard speed by means of which the three systems can be compared for evaluation purposes is that of the largest text size in Kb that could be displayed in the text boxes of NA, which equals 29 Kb, i.e., about 4,981 words. Therefore, a typical text of this size is prepared constituting a collection of the twelve sample texts in this study. This large

<table>
<thead>
<tr>
<th></th>
<th>Speed W/Sec</th>
<th>Speed %</th>
<th>Speed W/Sec</th>
<th>Speed %</th>
<th>Speed W/Sec</th>
<th>Speed %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological</td>
<td>351.5</td>
<td>40</td>
<td>351.5</td>
<td>40</td>
<td>175.75</td>
<td>20</td>
</tr>
<tr>
<td>Chemical</td>
<td>570</td>
<td>66</td>
<td>570</td>
<td>66</td>
<td>142.5</td>
<td>16.6</td>
</tr>
<tr>
<td>Computational</td>
<td>458</td>
<td>53</td>
<td>458</td>
<td>53</td>
<td>152.6</td>
<td>17.8</td>
</tr>
<tr>
<td>Financial</td>
<td>484</td>
<td>56</td>
<td>484</td>
<td>56</td>
<td>121</td>
<td>14.1</td>
</tr>
<tr>
<td>Legal</td>
<td>535.5</td>
<td>62</td>
<td>267.7</td>
<td>30</td>
<td>119</td>
<td>13.9</td>
</tr>
<tr>
<td>Medical</td>
<td>576</td>
<td>67</td>
<td>576</td>
<td>67</td>
<td>144</td>
<td>16.8</td>
</tr>
<tr>
<td>Military</td>
<td>567</td>
<td>66</td>
<td>567</td>
<td>66</td>
<td>81</td>
<td>9.4</td>
</tr>
<tr>
<td>Petroleum</td>
<td>* 855</td>
<td>100</td>
<td>* 855</td>
<td>100</td>
<td>213.75</td>
<td>25</td>
</tr>
<tr>
<td>Physical</td>
<td>630</td>
<td>73</td>
<td>630</td>
<td>73</td>
<td>105</td>
<td>12.2</td>
</tr>
<tr>
<td>Social</td>
<td>585</td>
<td>68</td>
<td>585</td>
<td>68</td>
<td>146.25</td>
<td>17.1</td>
</tr>
<tr>
<td>Commercial</td>
<td>220.5</td>
<td>25</td>
<td>110.25</td>
<td>12.8</td>
<td>73.5</td>
<td>8.5</td>
</tr>
<tr>
<td>Water Engin.</td>
<td>480</td>
<td>56</td>
<td>480</td>
<td>56</td>
<td>120</td>
<td>14</td>
</tr>
<tr>
<td>Average Values</td>
<td>526</td>
<td>61%</td>
<td>472</td>
<td>57%</td>
<td>133</td>
<td>15.5%</td>
</tr>
</tbody>
</table>

![Table (5.9). The Effect of Processor Speed on the Translation Speed.](v)

* The maximum speed that each speed of translation in the above table is compared with to obtain the evaluation speed percentage.
text is translated by the three participating systems. Unsurprisingly, GW is the fastest among the other systems in performing the translation of this text in 7 seconds merely. The results of MA and NA are 15 and 42 seconds, respectively. Table (5.10) below ranks the systems on the basis of these results in terms of the number of words translated per 1 second.

Table (5.10). The Fastest MT System in Translating Typical Text of 29 Kb.

<table>
<thead>
<tr>
<th>MT System</th>
<th>Speed of Translation (W/Sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GW</td>
<td>712</td>
</tr>
<tr>
<td>MA</td>
<td>332</td>
</tr>
<tr>
<td>NA</td>
<td>119</td>
</tr>
</tbody>
</table>

5.1.1.2.2 Adaptability Results

Results of the benchmark test are represented in Table (5.11) below, where the two evaluators have examined the extent to which each of the tested MT systems is adaptable to different operational settings. As it is an objective test, i.e., based on facts and direct observations, the results of the two raters are entirely identical and consistent.
Table (5.11). Average Values of Adaptability Based on the Checklist Results for the Arabic MT Systems.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value of MT System for the Two Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GW</td>
</tr>
<tr>
<td>Adaptable</td>
<td>100%</td>
</tr>
<tr>
<td>Unadaptable</td>
<td>-</td>
</tr>
</tbody>
</table>

It is worth stressing, here, that although GW and MA can only be installed on one PC at a time for security purposes, they both, in fact, can be registered on any PC of whatever features in case of direct on-line contact with the manufacturers. Moreover, since the two systems can run under any operating system, i.e., Windows 98/2000/ME/NT/XP, they are considered completely adaptable by the two evaluators and have been given a value of (4) in the checklist (See Appendix C), i.e., they are 100% adaptable. Further, NA system can easily be installed and operated on different PCs of various specifications. It should be noticed that the system tested in this study is not the original, but a copy of NA, so that it can be easily ported to different operational environments. However, this system cannot run under either Windows 2000 or XP. Thus, in the final evaluation, the raters have given it a value of (3) in the checklist, which constitutes an adaptability of 75%, i.e., a good adaptability.

5.1.1.2.3 Storage Results

In order to investigate the largest text size for translation by the Arabic systems in this study, general texts of various sizes in Kb have been tested (See 4.3.1.2.3). As a result, it is found that the largest size of text that NA
can display in the text boxes English and Arabic is 29 Kb merely. Since GW and MA are produced by the same company (ATA), the results point out to a largest text size of 40 Kb that both systems can display in their boxes, which exactly corresponds with the systems’ storage indicated by the manufacturers. This is considered typical system storage in this study by means of which the average values of this criterion for the participating systems is calculated using the method given in (3.3.2.3.8). It is also noticeable that even in cases where texts (exceeding the maximum storage of the system) are inserted as input data for translation, the system either displays the part of the text that corresponds with its maximum storage in its text boxes or, sometimes, it rejects the text entirely, the thing that results in closing the system.

As stated earlier in this study (See 2.11.4.8.1.5 and 2.11.4.8.3), both GW and MA have the option of translating very large documents regardless of their length by means of direct translation in the background. Accordingly, general texts of large sizes representing collections of various domains have been multi-translated (i.e., where the English text and its Arabic translation appear in one document paragraph-by-paragraph) by both systems. The results of these tests are manifested in Table (5.12) below. It is worth stressing, here, that while preparing these large texts, the process of addition of new words to increase the text size is done step-by-step. In other words, there is a limit of words addition indicated by the appearance of a message box to point out to unavailability of enough memory. To increase available memory, the researcher saved the text under evaluation with the new additions, and then, opened it again repeating the same task of adding new words to increase its size. This process, in fact, is time-consuming, but it is done only for experimental purposes.
Table (5.12). Direct Translation in the Background for GW and MA.

<table>
<thead>
<tr>
<th>Text Size Kb/Mb</th>
<th>Number of Words per Text</th>
<th>Time of Translation Hr:Min:Sec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GW</td>
<td>MA</td>
</tr>
<tr>
<td>282 Kb</td>
<td>58.292</td>
<td>00:2:5</td>
</tr>
<tr>
<td>1.45 Mb</td>
<td>233.165</td>
<td>00:10:00</td>
</tr>
<tr>
<td>2.91 Mb</td>
<td>466.329</td>
<td>00:20:24</td>
</tr>
<tr>
<td>5.84 Mb</td>
<td>1.165.823</td>
<td>00:53:46</td>
</tr>
<tr>
<td>23.3 Mb</td>
<td>5.362.786</td>
<td>4:06:00</td>
</tr>
</tbody>
</table>

In this respect, the user of these systems and even of NA has another option of splitting the original file into smaller ones and translating them separately and then he can merge the Arabic files (TTs) into one. Yet, this process needs much time and effort as well.

Taking the maximum storage of the systems and their extra-capacity of performing direct translation of large documents in the background into consideration, the average values of storage for the participating systems are calculated (See Table 5.13). On the basis of this, the storage capacity of both GW and MA is considered the highest in this evaluation by the two raters. This represents an average value of 100% for the highest systems’ performance in this respect. The average value of NA is less than that of GW and MA for its lower storage in comparison to the two systems and its inability to perform direct translation of large documents in the background. Thus, the average value of the system’s storage is 72.5% on the basis of the method of measurement adopted in the evaluation of this criterion (See 3.3.2.3.8), which is also good.
Table (5.13). Average Values of Storage Based on the Results of the Two Evaluators for the Arabic MT Systems.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value of MT System for the Two Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GW</td>
</tr>
<tr>
<td>High Storage</td>
<td>100%</td>
</tr>
<tr>
<td>Fair Storage</td>
<td>-</td>
</tr>
<tr>
<td>Low Storage</td>
<td>-</td>
</tr>
</tbody>
</table>

5.1.1.2.4 Flexibility Results

In the present study, the flexibility test mainly aims at finding out the capacity and possibility of adding new words and dictionaries to the participating systems. This involves lexical flexibility, merely, with respect to dictionary building and addition of new words, expressions or domain-specific terms to the system’s dictionaries.

Since there is an option of free Internet updates for the most recent versions of GW and MA concerning the systems’ dictionaries (i.e., general and special-domain dictionaries), this is considered an indication of high system’s extensibility. Moreover, with the exception of GW, both MA and NA allow the user to create and add new dictionaries to the systems. Thus, evaluation of this criterion is based on the tests and discussion presented in (4.3.1.2.4). In this regard, the researcher has designed a simple table below (Table 5.14) where the capacity of each Arabic MT system is displayed in terms of lexical flexibility.

Table (5.14). Lexical Flexibility of Arabic MT Systems.

<table>
<thead>
<tr>
<th>System’s Capacity</th>
<th>GW</th>
<th>MA</th>
<th>NA</th>
</tr>
</thead>
</table>
Therefore, the two evaluators have based their rating on the results of this table and the features in the checklist (See Appendix C). Accordingly, the average value of this criterion for each of the participating systems is calculated. Table (5.15) reflects the results of the final evaluation. Among the other MT systems, MA has been given score (2) in the checklist (i.e., very good flexibility) with the highest average value of 100% for lexical flexibility as the system allows the user to update and add new dictionaries. As there is no option of dictionaries building with GW, the two raters have assigned the system score (1) in the checklist (i.e., limited flexibility) for the addition of new words, expressions and terminologies when updating its general and special-domain dictionaries. This rate constitutes an average value of 50% for this system. Besides, NA has also been given a score of (1) in the checklist with an average value of 50% for its flexibility in dictionary building merely.

Table (5.15). Average Values of Flexibility Based on the Checklist Results for the Arabic MT Systems.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value of MT System for the Two Evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GW</td>
</tr>
<tr>
<td>Flexible</td>
<td>50%</td>
</tr>
<tr>
<td>Inflexible</td>
<td>-</td>
</tr>
</tbody>
</table>
However, the process of dictionary building is not an easy task with NA, since the system available in the local market is a copy of the original. Thus, the researcher has faced certain problems in designing and operating the system’s dictionaries, which will be discussed later in this chapter.

![Figure (5.8). Average Values of Non-Functional Criteria Showing the Range of Arabic MT Systems.](image)

On the whole, Figure (5.8) above clearly shows the ranking of the three Arabic MT systems with respect to each of the evaluated non-functional criteria. In general, the average values of MA are the highest with top results of 100% except for the speed of translation where its value is lower than that of GW. Nevertheless, this does not affect ranking the system with the best non-functional performance. Although its average value for speed is higher than that of MA and its average values for adaptability and storage are equal to these of MA, GW comes second in rank after MA regarding the overall non-functional performance. This result is mainly due to its fair flexibility. Further, the results of NA are the worst among these systems as the average values it has obtained rank the system in the lowest position for non-
functional performance, except with flexibility as its value is the same as that of GW.

5.1.2 Evaluation Results of GSQ

Having obtained the average values of the functional and non-functional criteria for each of the participating MT systems, the overall GSQ is calculated for purposes of final systems’ assessment and comparison in terms of two levels: a.) criteria type and b.) text type.

5.1.2.1 Systems’ Quality Assessment and Comparison for Criteria

Based on the formula and discussion presented in (3.3.2.4), the overall average value of software quality for each of the evaluated MT systems is estimated. This is obtained from the average values of the functional and non-functional criteria in this study. In fact, the MT systems are judged and compared for their total performance and the evaluation aims at investigating which system’s components work adequately. Thus, the results of the sub-criteria, the mother-node criteria and the GSQ are presented below with respect to their average values and values of relative importance:

(1) According to the results manifested in Table (5.16) which are also graphically represented in Figure (5.9), MA has the best system’s total performance in comparison with GW and NA. It has obtained the highest percentage value of 58 for GSQ. This indicates that with respect to all the functional and non-functional (except for speed of translation) quality characteristics in this study, it gives the best results. Nevertheless, the system’s performance is not an optimized one especially with the functional criteria. In other words, the relative amount of degradation in translation quality is due to certain inadequacies in the contributing components of the
system. However, both GW and NA have nearly gained identical results, which are, in fact, not very much encouraging and below average\(^{(2)}\). Their total percentage values in the final evaluation of GSQ are 44 and 45, respectively. Thus, it is believed that there is an elemental need to make more efforts to develop and improve the major components of these systems.

Table (5.16). General Software Quality Based on Functional and Non-Functional Criteria.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Machine Translation System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Golden Al-Wafi</td>
</tr>
<tr>
<td>Readability</td>
<td>7.76</td>
</tr>
<tr>
<td>Fidelity</td>
<td>6.27</td>
</tr>
<tr>
<td>Syntax</td>
<td>6.36</td>
</tr>
<tr>
<td>Terminology</td>
<td>3.62</td>
</tr>
<tr>
<td>Morphology</td>
<td>4.72</td>
</tr>
<tr>
<td>Speed</td>
<td>3.0</td>
</tr>
<tr>
<td>Adaptability</td>
<td>5.0</td>
</tr>
<tr>
<td>Storage</td>
<td>5.0</td>
</tr>
<tr>
<td>Flexibility</td>
<td>2.5</td>
</tr>
<tr>
<td>Total Quality %</td>
<td><strong>44.0</strong></td>
</tr>
</tbody>
</table>

In spite of this, the percentage values of the Arabic systems in this study, regarding their total quality performance, rank MA at the top position followed by NA and only in the lowest position comes GW.
Below is the graphic representation of the total quality results that appear in the above table, where it seems obvious that MA has the best quality performance among the other Arabic MT systems evaluated with the total percentage value of 58.

![Figure (5.9). The Percentages of the Total Quality for the Three Arabic MT Systems.](image)

(2) Within the framework of functionality, results of the total performance of the mother-node criteria are also calculated in terms of the value of relative importance (See Table 5.17). Here, MA has got the highest values for ‘suitability’, ‘accuracy’ and ‘wellformedness’. These are 11.11, 16.07 and 13.33, respectively. As for NA, the quality of the translated output in this respect is better than that of GW, but worse than that of MA. The values for ‘suitability’, ‘accuracy’ and wellformedness’ of this system are 8.44, 14.04 and 12.18 showing no great difference from these of MA. Further, the results of GW for the same sequence of criteria are 7.76, 9.89 and 11.08 ranking the system in the lowest position in comparison with the first two. It is worth
stressing, here, that the functional performance of all the three Arabic MT systems is either fair (i.e., average) or bad owing to certain defects in their grammatical and semantic components.

Table (5.17). Average Values of Mother-Node Criteria in Terms of Systems’ Functional Performance.

<table>
<thead>
<tr>
<th>MT System</th>
<th>Suitability 16%</th>
<th>Accuracy 32%</th>
<th>Well-Formedness 32%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GW</td>
<td>7.76</td>
<td>9.89</td>
<td>11.08</td>
</tr>
<tr>
<td>MA</td>
<td>11.11</td>
<td>16.07</td>
<td>13.33</td>
</tr>
<tr>
<td>NA</td>
<td>8.44</td>
<td>14.04</td>
<td>12.18</td>
</tr>
</tbody>
</table>

(3) In the final stage of classification and presentation of the evaluation results for GSQ, the Arabic systems are compared for their best functional and non-functional total performance. As mentioned earlier in this study (See 3.4.2.4), the functional criteria are favored for their direct effect on the translation quality of the output. In general, the values for both functional and non-functional total performance of MA are the best among these systems. It has obtained a value of 40.15 for the former and 17.85 for the latter (See Table 5.18). This points out to fair quality of the MT output and high computational capabilities of the system. On the other hand, although the values for GSQ of GW and NA are almost identical, NA has a better functional performance than GW, i.e., 34.35 in comparison with 28.5; while

Table (5.18). GSQ Based on Functional and Non-Functional Total Performance.

<table>
<thead>
<tr>
<th>MT System</th>
<th>Functional Total</th>
<th>Non-Functional Total</th>
<th>GSQ 100%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
GW is better in the computational capabilities than NA. The values of these results are 15.5 and 10.65, respectively. Thus, the quality of the MT output produced by NA is better than that of GW which is faster in performing the translation and capable of translating larger texts and documents.

### 5.1.2.2 Systems’ Quality Assessment and Comparison for Text Types

The GSQ of each of the evaluated systems is also calculated in terms of text types, i.e., a total quality obtained from the average values of all the text types translated by an MT system. Here, it is also intended to see in which domain(s) an Arabic system produces the best MT output and in which domain(s) it translates worst. Moreover, the evaluation aims at investigating whether each system is better in translating scientific or humanity texts (e.g., social, legal and commercial) and which system is the best in this respect. Further, although the manufactures of the Arabic MT systems have claimed that their systems are made to translate any text type of whatever domain, they have also suggested certain fields where these systems can translate best. It is worth stressing, here, that most of these text types are taken into consideration in the present work, partly to see if the system’s performance corresponds with the recommendations of its manufactures. So, with reference to Figure (5.10) below, the final evaluation of the participating systems gives the following results:

<table>
<thead>
<tr>
<th></th>
<th>performance 80%</th>
<th>Performance 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GW</td>
<td>28.5</td>
<td>15.5</td>
</tr>
<tr>
<td>MA</td>
<td>40.15</td>
<td>17.85</td>
</tr>
<tr>
<td>NA</td>
<td>34.35</td>
<td>10.65</td>
</tr>
</tbody>
</table>
Figure (5.10). The Variation of Total Quality for the Three MT Systems Based on Text Types.

(4) With respect to GW system, the total quality it has obtained in terms of text types is 44% which is below average. In other words, for the majority of text types translated by this system, the range of the total quality performance is from 30% to 46% except for three text types (computational, social and commercial) where the system has performed best obtaining the average values of 54%, 56% and 60%, respectively. Although ATA Company has recommended GW for best performance in the scientific fields, it has been proved that the system is better in translating humanity and computational texts merely. Thus, showing certain defects in the grammatical, lexical and semantic components, GW can produce average MT output with fair fluency and accuracy with computational, social and commercial texts. Otherwise, the quality of the MT output of this system is poor.

(5) The results of MA are more encouraging than these of GW getting the best total quality performance of 58% in comparison with the other systems. In general, the average values of the texts’ quality range from 52% to 69% which are either average or above average. The system has the best performance with the computational, legal, water engineering and petroleum
texts. The percentage values of these results are 69, 64, 62 and 60, respectively. In spite of its best performance among the other systems and obtaining above average results regarding most of the domains it has been suggested to translate, MA requires serious efforts to develop and improve the system’s components especially the syntactic and the morphological ones which badly affect the fluency and accuracy of its output. Based on the results of the present evaluation, the system shows the highest performance with the scientific texts, although, the average values of all the texts reflect no significant differences. In brief, MA can provide fast, readable and accurate MT output of average quality in most cases.

(6) The system that comes second in rank after MA with an almost identical total quality performance with GW (i.e., 45%) is NA. This result also reflects a total texts’ quality below average. For the majority of texts translated by this system, the range of values is from 34% to 49%. It is only with the social, legal and commercial text types, the system has the best performance, which is above average. The percentages of these results are 53, 52 and 50, respectively. Although the results presented in Table (5.18) in (5.1.2.1) indicate that NA has a better functional performance than GW, the quality of the MT output of this system is also poor. This is attributed to the same reasons of producing fair or bad MT output by the other systems. Like GW, NA is better in translating humanity texts than scientific ones. It has a low performance with respect to most of the text types recommended by its manufactures and even with its best translations, the result is slightly above average in most cases. All in all, this system is inadequate in performing the translation of military and scientific texts.

Thus, in the present evaluation, it has been proved that the functionality of MA is superior to that of the other MT systems. Nevertheless, it is still
notably within average and only enables the system to give a draft translation of fair quality. In addition, functionality has a number of sub-nodes, i.e., suitability, accuracy and well-formedness. It has also been confirmed that MA has the best performance among the other systems for all these sub-characteristics. However, the system points out to fair accuracy with respect to its conformity to the specifications laid down for it. Its suitability is somehow above average as the texts it translates are, in general, understandable and have clear meaning only after some reflection, and it is more suitable for scientific texts. The other MT systems (i.e., GW and NA) are more suitable for humanity texts than scientific ones and their accuracy is below average. For the three systems, the results of evaluating syntax and morphology reflect marked weakness.

Furthermore, results of the non-functional criteria indicate that MA has again the highest performance for it gives fast translation of large texts and documents in few seconds only, and it is not a closed system, i.e., it has very good portability as it can be easily replaced by new version as a result of system improvement over time. Besides, it is adaptable to any PC of whatever specifications and runs under any version of the operating system (i.e., Windows). As for the other systems, GW has been proved to be faster than MA, but its flexibility is limited to dictionary updating only. Moreover, NA has been found inferior to the other systems in this respect as it is slower than GW and MA and is inflexible to the addition of new words. It also cannot translate large documents and is inadaptable to the very recent versions of the operating system.

Finally, before drawing conclusions from the present evaluation, it is important to sum up some of the sources of problems for inaccurate,
unsuitable and ill-formed MT output of the three Arabic systems that can be used for personal and on-line translation services.

5.2 Problems of MT Quality

The preceding sections have given the results of the experimental analysis and evaluation of the functional and non-functional criteria, as well as systems’ quality and total performance for criteria types and text types as employed in the data. The next few sections will discuss these results, outlining with exemplification from the MT output of the three evaluated systems the type of MT problems, their extent and causes.

For convenience, these quality problems can be classified into three types: cognitive problems, linguistic problems and operational problems. The cognitive problems concern fluency and understandability of the MT output (i.e., text readability and clarity of meaning), and its syntactic and semantic fidelity to the ST. The linguistic problems involve inexact translation of domain-specific terms, and inappropriate and misuse of syntactic constructions and morphological forms in the MT output of the TL. Operational problems are related to speed of translation measurement and systems’ dictionaries making. It must be noted that these problems overlap to a large extent and they contribute in their combination to inefficient and poor quality of MT output.

Because of limitation of space, illustrative exemplification of each type of problems will only be given where necessary and will be restricted to one or two excerpts culled from the data. Each example will be introduced with the tag “Ex.” to denote “excerpt” or “extract”. An excerpt representing a certain problem will show the SL sentences (ST), the reference translation
(HT) and one example from the data of each of the tested MT systems (TT). Not only extracts manifesting problems are given, but these reflecting correct and adequate translation (i.e., obtaining high scores in the evaluation scale) will also be referred to.

5.2.1 Cognitive Problems

Various types of problems have been identified in the data. These can be considered as direct causes of unintelligible and inaccurate MT output. Likewise, they virtually result in complete distortion of meaning and unfaithful conveyance of the content of the ST. Indeed, linguistic phenomena are particularly troublesome for such problems. Many grammatical and word usage problems can be identified in this respect, but it is preferable to discuss them in detail in (5.2.2). The main factors which give rise to incomprehensible and unclear meaning of the Arabic MT output, and affect its fidelity to the ST are illustrated below:

5.2.1.1 Grotesque Word Order and Choice

Pertaining to fluency and understandability of TL sentences, syntactic correctness, organization and coherence, and stylistic appropriateness are required. As already noted, a considerable number of sentences translated by the Arabic MT systems dramatically reflect complete alteration and distortion of meaning. The idea can be vaguely apprehended. Word choice, syntactic arrangements and alternative expressions are generally bizarre. Besides, there are critical words untranslated. These sentences are, in fact, unintelligible and they tend to read like nonsense. It appears that no amount of study and reflection would reveal the thought of the sentence. This can be seen in the following examples:

Ex.1
ST: When subsequently, each (double-structured) member of the homologous pair splits longitudinally, one or more transverse breaks occur in the chromatids and an interchange of chromatid segments between two homologous chromosomes occurs (Fig. 1-2 C).

(From a medical text)

HT: وبالتالي حين ينشطر طوليا كل عضو في الزوجين المتماثلين (مزدوج التركيب) يحدث انكسار مستعرض أو أكثر في الكروماتيدات ولهذا يحدث تبادل للجزء الكراماتيدية بين الكروموسومين المتماثلين (شكل 1-2 ج).

GW: عندما، بعد ذلك، كل (ضعف نظم) ينشطر عضو الزوج المشابه طوليا، يحدث واحد أو أجازات مستعرضة أكثر في chromatid وتبادل قطع Chromatids بين كروموسوم مشابه يحدثان (تينه.1-2 سي).

MA: عندما، بالتالي، كل (ضعف بني) عضو تزدوج ينشطر طوليا المشابه، واحد أو مستعرضة انكسارا يحدث الأكثر من chromatid وتبادل من chromatids الأجزاء بين كروموسوم مشابه يحدثان (تينه. 2-1 سي).

NA: عندما، بالتالي، ينشق كل (ضعف المنفي) العضو للزوج المتماثلة طوليا، أحد أو تقع أكثر أكثر الكسور مستعرضة أو في الكروماتيد وتبادل شرائح كروماتيد بين اثنان الكروموسوم المتماثلة تقع (شكل 1-2 سي).

Ex. 2
ST: One member of the pair (the dextrorotatory, or D, isomer) rotates plane-polarized light to the right, while the other (the levorotatory, or L, isomer) rotates it an equal amount to the left.

(From a chemical text)

HT: يدير احد أعضاء زوجي الأشياط الجزئية (وهو الذي يدور باتجاه عقرب الساعة, أو الشبيه د) الضوء المستقطب استوائيا إلى اليمين، بينما يدير الآخر (وهو الشبيه الذي يدور عكس عقرب الساعة، أو الشبيه ل) كمية الضوء ذاتها إلى اليسار.
The above examples show a combination of grammatical and word usage problems in the MT output of the three Arabic systems. Needless to say, the translations done by MA and NA are better than those of GW system which is mainly due to the effect of the user dictionary. Nevertheless, they are still examples of poor rendition. The translation style of GW seems almost like that of MA since the two systems are produced by the same company. They use the same MT modules, but GW has a less extensive lexicon. In general, what affects the clarity of these examples can be attributed to incorrect grammatical forms and problems in the relationships between the sentence elements. The common feature concerns preservation of the SL word order in the TL text. In other words, the texts are translated literally, word-for-word following the word order of the SL. Therefore, no dependencies between words are created. In addition, untranslated words, poor style and unhappy word choice are also noticeable. More examples showing complete alteration and distortion of meaning are presented below:

Ex. 3
ST: Other disbursements–outlays for property, plant, and equipment and long-term investments.

(From a financial text)

HT: مدفعات أخرى – هو أي مدفعات لأجل شراء العقارات و المباني و التجهيزات و الاستثمارات الطويلة الأجل.

GW: إنفاق أخر – تفقات للملكية، نباتات و أجهزة و طويل تعبير الاستثمارات.
ST: Other memory-addressing modes allow virtually image zoom, as well as, scroll (vertical shifts) and pan (horizontal shift)

(From a computational text)

HT: وعملياً تسمح الأساليب الأخرى لعنونة الذاكرة بتقريب وإبعاد آني للصورة بالإضافة إلى الإزاحة العمودية والأفقية.

GW: الذاكرة الأخرى التي تخطب الألماط تسمح لعملياً ازيز صوره، بالإضافة إلى اللفيفه (نوبات عمودية) ومقلاة (تغير أفقي).

Ex. 5

ST: Given the horrifying effects of such conflicts, enhancing political rights and inclusion should be a top priority for maintaining national integrity and peace.

(From a social science text)

HT: بالنظر إلى التأثيرات المروعة لمثل هذه النزاعات ينبغي أن يكون تعزيز الحقوق والتضمين السياسي في قمة الأسبقيات من أجل الحفاظ على الوحدة الوطنية والسلم.

NA: معطى مؤثرات مثل هذه النزاعات المروعة، ينبغي على الدقة السياسة وينبغي على التضمين أن يكون أولى الأولويات لوحدة مواطن الحفاظ والسلام.

Ex. 6

ST: Other such improvements include the provision of lighter and longer – range laser designators, and light all-terrain vehicles and trucks that offer higher mobility and less detectability than systems like the HMMWV.

(From a military text)

HT: وتشمل التحسينات الأخرى اللازمة تأمين منظمات تخصصي وتمييز الأهداف الليزرية اخف وزناً وأطول مدى، وعروض خفيفة ملائمة لجميع أنواع الأراضي وشاحنات توفر قابلية تحرك أعلى وقابلية اكتشاف أقل من منظمات هموف وهي شاحنة ذات قابلية حركة ومناورة عالية.
Totally clear, the English-into-Arabic translations performed by the three systems are quite disappointing. They still have serious contextual reasoning problems with getting the word order, the correct word forms, the grammatical categories and choosing the right word or expression out of a number of alternatives. Moreover, they either translate domain-specific terms by their most common sense, i.e., out of context or do not translate them at all. As a result, the sentences cannot be understood and no amount of effort will produce any meaning.

5.2.1.2 Understandable Clumsy Translation
Here, the sentence masquerades as intelligible, but actually it is more unintelligible than intelligible. The general idea is somewhat there, but it is hard to read due to awkward translation. In other words, problems in grammar, word usage and/or style and similar phenomena are present, but constitute mainly 'noise' through which the main idea is still perceptible. The following examples illustrate this point:

Ex.9

ST: Combined sewers tend to get silted and fouled during dry-weather flow unless they are laid at steep gradients or have cunettes (small longitudinal channels or troughs in the bottom of a large sewer used to concentrate the flow and increase the velocities for low flows).

(From a water engineering text)

HT: ويتكون في المجاري المشتركة الغرين والقاذورات خلال جريان مياه الفضلات في المناخ الجاف ما لم يكن فيها احترادات حادة (أو قنوات طويلة صغرى أو أخاذيد في قفر المجاري الكبير تستعمل لتركز وزيادة سرعة الجريان البطيء.

GW: تهتم بالبالوعات المشتركة بأن تصبح مغرين ومفسدة أثناء تدفق طقس جاف ما لم هم يضعون في الميول الحادة أولهم cunettes (قنوات أو منخفضات طويلة صغيرة في قاع بالوعة كبيرة تستعمل لتركز التدفق وتزيد السرعة للتدفق المنخفض).

MA: يهتم المجاري المشتركة بالحصول على غرين وقاذورات خلال طقس جاف جريان ما لمهم يضعون على حادة احترادات أو عندهم قنوات (طويلة قنوات صغرى أو أخاذيد في قفر مجرى كبير يستعمل لتركز جريان ويزيد سرعة للتدفق المنخفض).

NA: تميل مجاري مشتركة إلى أن تحتصل المغرينه وتنثوث أثناء صرف ه - تدفق الطقس ما لم يطرحون على الأرض بقوة في منحدرات أو عندهم قنوات (قنوات طويله صغرى أو كان أخاذيد في قفر مجرى كبير يركز التدفق ويزيد السرعة لتدفقات منخفضة).

Ex.10

ST: The fraction of the pore space occupied by water increases with depth in the transitional zone so that the base of the transitional zone is delineated by completely water-saturated pore space.
For the three Arabic MT systems, the examples given, here, contain many grammatical and word usage problems, and meaning can only be guessed after careful study, if at all. To clarify the meaning of these sentences, too many revisions would be needed.

5.2.1.3 Unknown and Mistranslated Critical Words

Sometimes the sentence could be generally clear and intelligible, although style and syntactical arrangements are somewhat poor. However, what badly affects its readability and clarity of meaning is the untranslation or mistranslation of critical words. Post editing could leave this in nearly acceptable form. This problem is manifested in the examples below:

Ex. 11

ST: Without the specialized flagellates (a "cluster" of species of the order Hyper-mastigina), many species of termites are unable to digest the wood they ingest, as shown by the fact that they starve to death when the flagellates are experimentally removed.

(From a biological text)
In the above examples, the translation of GW is poorer than that of the other systems. Neither MA nor NA produces good translations, although the gist can be extracted. The positive effect of the user dictionary of these two systems is apparent in their output too. In spite of certain grammatical and stylistic infelicities in these sentences, the untranslation and mistranslation of some critical words is the direct source of readability problems here.

Irrespective of word order and other grammatical and stylistic phenomena, the mistranslation of 'flagellates', 'cluster', 'species', 'order' and 'Hyper', in addition to the untranslation of 'mastigina' in the MT output produced by GW have seriously distorted meaning. Similarly, the misleading interpretation of the critical term 'flagellates' into 'يجلد' instead of 'سوطيات' by MA system has a bad effect on the understandability of the sentence. Further, the absence of the rendition of the subject compliment 'unable to digest the wood' (i.e., تعجز عن هضم الخشب) and the unhappy translation of other words in the MT sentence of NA system have affected the clarity of meaning.

5.2.1.4 Partially Correct Translation

Another recognizable phenomenon is that some sentences translated by the Arabic MT systems are partially intelligible as only certain parts of each have correctly been rendered. This has a
negative effect on the overall readability and clarity of their meaning. The following examples illustrate this point:

Ex. 12
ST: In either case, the user provides the computer with input information through a keyboard, which resembles an ordinary typewriter.

(From a computational text)

HT: وفي كلتا الحالتين يزود المستخدم الحاسب بالعلومات الداخلية من خلال لوحة المفاتيح التي تشبه آلها كاتبه عاديه.

GW: في اما يفتح المستعمل يزود الحاسوب مع inputinformation من خلال لوحة مفاتيح التي تشبه آلها كاتبه عاديه.

Ex. 13
ST: If both carbons of the double bond are also joined by single bonds to 2 different substituents, the structure may be drawn in 2 forms that are cis-trans isomers.

(From a chemical text)

HT: وإذا كانت ذرتا الكربون للأصية المزدوجة ترتبطان أيضاً بأواصر مفردة مع بديلين مختلفين يمكن رسم التركيب على هيئة شكلين شبيهين جزيئيين مقرونين – مفروعين.

MA: إذا كلا كاربون رابطه مزدوجة يربط أيضاً من قبل الفرد أواصر إلى 2 substituents مختلف، التركيب قد يرسم في صورتي ذلك الأشباه الجزيئية المقرنة – المفرقة.

Ex. 14
ST: The most obvious of these gaps is that Arab women remain marginalized and under-utilized in all arenas, notably in terms of their economic, intellectual and leadership potential.

(From a social science text)

HT: ومن ابرز هذه الفجوات بقاء النساء العربيات على الهمش و عدم استثمار قدراتهم في جميع الميادين لاسيما قدراتهم الاقتصادية والفكرية والقtemperature.

MA: وإذا كلا كاربون رابطه مزدوجة يربط أيضاً من قبل الفرد أواصر إلى 2 substituents مختلف، التركيب قد يرسم في صورتي ذلك الأشباه الجزيئية المقرنة – المفرقة.
Obviously, the last parts of the sentences translated by GW and MA and the first part of the sentence translated by NA represent correct and almost human like rendition. Nevertheless, these sentences are not examples of complete fluency and understandability for the awkward translation of the other part of each. It is also worth stressing, here, that the longer the sentence translated by these MT systems, the more unintelligible its meaning would be.

As stated above, the MT outputs of the three Arabic systems reflect many grammatical and word usage deficiencies. For these drawbacks, translations done by these systems have shown to produce texts which are not faithful or accurate with respect to the ST in addition to their being quite unintelligible. Thus, problems affecting the faithful conveyance of the content of the input sentence to the output sentence (as observed in the data) are discussed under two headings:

5.2.1.5 Adherence to SL Structures

One of the most distinctive features of texts produced by MT systems is their unnatural literalness. In general, they adhere too closely to the structures of STs. Although Newmark (1991) considers literalness to be desirable in literary and authoritative texts, as long as the result is in the appropriate style, the aim in technical/scientific translation is generally to produce texts which read as if they were originally written in the TL. Indeed, it is quite evident that MT systems do not achieve this goal, especially if the translation is performed between unrelated languages (e.g., English and Arabic).

Accordingly, the principal reason for the absence of fidelity in the output of the systems considered is due to complexities of structural transfer and generation of Arabic grammatical constructions. In fact, structural changes are so common when translating from one language into another that the 'low-level' ordering of basic elements-nouns, verbs, adverbs and adjectives should be expected from any MT system. In Arabic output, for instance, the syntactic dependencies and relationships between the sentence elements should be in the correct Arabic order (e.g., الاتجاهات الأخرى لعنونة الذاكرة instead of الذكاءات الأخرى - يعنون اسلابياً) , a word-for-word translation done by NA system of the English construction (Other memory-addressing modes). Not only order is of significance, here, but all the other factors like word forms and inflections.
and the grammatical categories and arrangements should be considered with reference to the TL grammatical rules in addition to their being faithful to the ST (e.g., passive or plural forms etc.). These must be regarded as the minimum requirements, and any MT system which fails in this respect must be suspected of deficiencies elsewhere, probably of a graver nature.

To avoid repetition, no detailed identification of the grammatical phenomenon causing infidelity is given here (See 5.2.2 below). These are only to be referred to in general and within the limits of the following examples:

Ex. 15

ST: Total cash needed disbursements, plus the minimum ending cash balance desired.

(From a financial text)

HT: ويشمل إجمالي النقد المطلوب إجمالي المدفوعات، زائداً الحد الأدنى لرصيد النقد المطلوب في نهاية المدة.

GW: النقد الكلي إحتاج يتضمن مجموعاً الإتفاق، زائدًا رصيد بصندوق الإنهاء الأدنى رغب.

Apparently, the content of the input sentence is not adequately conveyed in the output sentence translated by GW system. This is mainly attributed to the word-for-word rendition of the sentence violating the syntactic relationships between words and phrases, and the appropriate word positions in the TL.

Ex. 16

SL: I have a fixed abode to which may be forwarded all the judicial and executory documents.

(From a legal text)
In this example, the word choice problem has been overcome by the user dictionary. However, the unfaithful conveyance of ST information and distortion of meaning in the MT output of MA is obviously due to literal rendition and bad syntactic parsing.

Ex. 17

ST: It cannot take off in severe rain, snow, ice, or fog conditions; its imagery lacks the definition to find and characterize some types of targets: it is a slow flier (90 MPH) that operates best at 10,000 feet which puts it within range of many forms of light anti-aircraft defense and which has led to losses in Afghanistan and Iraq; it has awkward control systems and ergonomics; and each unit (four planes and a ground station) costs about $25 million.

(From a military text)
Despite some good word choice which is chiefly the result of the user dictionary, the content of the ST is not adequately conveyed in the above MT output of NA system. A comparison with the ST and HT points to a combination of syntactic problems which cause infidelity and uncleanness of meaning as well. These concern inappropriate grammatical forms and relationships between words and expressions; misuse of gender, number, case and definite article; wrong choice of word class items and poor style, in general.

5.2.1.6 Loss and Distortion of Information
In addition to problems of syntactic fidelity, there exists a kind of lack of semantic faithfulness in a number of sentences in the data. This indicates that the content of the input sentence is not conveyed in the MT output sentence at all. In other words, the output is not a proper sentence. This is mainly due to: a.) loss of information (e.g., where words or even whole sentences are not translated or clauses or phrases are missing), b.) interference (i.e., noise) where there are words or expressions added by the
system and c.) a distortion of form—a combination of loss and interference when some words are badly translated or do not have an equivalent in the TL. The sentences offered in (5.2.1.1 above) serve to illustrate this point. Further examples are given below:

Ex. 18

ST: Two types of stereoisomers are commonly encountered in biochemistry: mirror image isomers (enantiomers and cis-trans isomers (geometric isomers).

(From a chemical text)

HT: ونجد عادة نوعين من الأشباه الجزيئية أنمجسمه في الكيمياء الحياتية هما الأشكال المتقابلة المراوية والأشباه الجزيئية المقربة – المفرقة (الأشباه الجزيئية الهندسية).

GW: اثنان من أنواع stereoisomers يصادف عموماً في الكيمياء الحيوية: ألصورة المطابقة isomers (isomers (enauti omers

Ex. 19

ST: We regret to inform you that we could not supply you with more than the fifty machines model ... because of unexpected hold-up in production.

(From a commercial text)

HT: نأسف أن نخبركم إننا لم نستطيع أن نرسل لكم أكثر من المكان الخمسين موديل ... بسبب
Ex. 20

ST: In this case the symbionts live inside the body of the host, but even intimate interdependence may develop with the microorganism partners living outside the body of the animal host, and such associations may actually represent a more advanced stage in the evolution of mutualism (less chance that the relationships might revert to parasitism!).

(From a biological text)
they do not exist in the general dictionary of the system. As for MA, very few instances of untranslated words or expressions are found in the data of this system due to the effect of the user dictionary. However, in the above example of MA, there is missing information for the untranslation of the word 'unexpected'. Regarding the MT sentence translated by NA, the whole noun phrase 'a more advanced stage in the evolution of mutualism' (i.e., مرحلة أكثر تقدمًا في إطار تطور تبادلية المنفعة) is missing. This results in loss of information affecting the clarity of meaning. It is worth stressing, here, that two other sentences in this text have not been translated at all and the same thing occurs with other text types translated by the same system. By following a word-for-word literal translation, it is hard to see any instance of words or expressions added by the system in the data of GW. Noise in the MT output of this system is caused by other phenomena (See 5.2.1.2 above). As for both MA and NA systems, very few examples of additional or repeated words and expressions, affecting the fidelity of the MT output to the input SL sentence, are noticeable in the data of these systems. Nevertheless, some examples that illustrate this point are observed in the MT sentences of the two systems (this phenomena is more apparent in almost all the text types translated by NA system), as seen below:

Ex. 21

ST: Many common optical devices contain not only mirrors and prisms having flat polished surfaces but lenses having spherical surfaces with a wide range of curvatures.

(From a physical text)
Ex. 22
ST: The centromere regions of the homologous chromosomes do not pair.

(From a medical text)

Ex. 23
ST: Carbon plays a central role in biochemistry ....

(From a chemical text)
Interactive computing is carried out either with small desk-top computer or with a computer terminal.

(From a computational text)

Soon afterwards, each chromosome undergoes ...

(From a medical text)

In analogy to... (See Chap. 3), a plane perpendicular ...

(From a physical text)
Ex. 27
ST: We are therefore able to offer a wide range of these toys at low prices. We can offer most of the items at a special discount rate of 50% off catalogue prices.

(From a commercial text)

HT: لذلك نستطيع أن نعرض عليكم تشكيلة واسعة من هذه الألعاب بأسعار زهيدة. ونستطيع أن نقدم معظم هذه الأنواع بخصم خاص بمعدل 50% عن أسعار الفرس.

NA: نحن نستطيعون لذلك من نموذج لعرض سلسلة عريضة من هذه الألعاب في أسعار منخفضة. نقدر أن نموذجًا لعرض أكثر المفردات في سعر الخصم الخاص من 50% من أسعار فرس.

Finally, a distortion of form and misapprehensions about general meaning, make an MT sentence not informative. This is attributed to a combination of untranslated individual words and some interference that result from the addition of irrelevant words or incorrect rendition of other words (i.e., deformed words) in the same sentence. All these problems cause unfaithful and inaccurate conveyance of information between the ST and the TT – the MT output of the systems. This is illustrated in most of the examples offered in (5.2.1.1 and 5.2.1.3 above). Other examples representing this case are:

Ex. 28
ST: Mineral cycling as well as food production is enhanced by symbiosis between microorganisms and plants.

(From a biological text)
The nomenclature of mirror image isomers was historically based on the observation that some pairs of isomers rotate the plane of plane polarized light.

(From a chemical text)
Other such improvements include the provision of lighter and longer–range designators, and light all-terrain vehicles and trucks that offer higher mobility … .

(From a military text)

Ex. 31


(From a legal text)

A look at the above excerpts demonstrates serious system’s deficiency to maintain precision and correctness in the MT output. In most of these
sentences, infidelity to the input SL sentence is due to untranslated words (i.e., either missing or kept in their SL form), bad rendition and inappropriate word choice irrespective of context and inadequate word-forward translation.

In spite of the above-mentioned cognitive problems reflecting unfaithful and unclear rendition of ST in the MT output, a number of sentences have been recognized in the data of the three Arabic MT systems indicating complete intelligibility and clarity of meaning where grammar, word usage and style are almost appropriate. Moreover, the content of the input sentence is faithfully conveyed to the output sentence. In sum, the translated sentence is readable and accurate and only very slight post-editing is needed. The following excerpts are typical examples:

Ex. 32
ST: In subsequent stages the members migrate to the opposite poles of the cell (Fig. 1-2E).

(From a medical text)

HT: وفي مراحل تالية، يهاجر الأعضاء إلى القطبين المتضادين للخلية (الشكل 1-2 د).

GW: وفي المراحل اللاحقة، يهاجر الأعضاء إلى الاقطاب المعاكس للخلية (تبنية 1-2 إ). 

Ex. 33
ST: However, significant gaps remain in the mobilization of human capabilities in Arab countries.

(From a social science text)
Ex. 34
ST: Sewerage systems are of two types: combined and separate.

(From a water engineering text)

Ex. 35
ST: The function of the display module is to read an image memory, convert the stored digital information into an analog video signal, and output this signal to a TV monitor or other video device.

(From a computational text)
ST: These factors block the full integration of women into the economic and intellectual life of their countries.

(From a social science text)

HT: وتمنع هذه العوامل الإدماج الناتج للنساء في الحياة الاقتصادية والفكرية لأقطارهم.

MA: تمنع هذه العوامل الإدماج للنساء إلى الحياة الاقتصادية والثقافية من أقطارهم.

Ex. 37

ST: Mineral cycling as well as food production is enhanced by symbiosis between microorganisms and plants.

(From a biological text)

HT: يعزز التكافل بين الكائنات المجهريّة والنباتات التدوير المعدني وإنتاج الغذاء.

MA: تدوير معدني بالإضافة إلى إنتاج الأغذية معزز من قبل تكافل بين كائنات مجهريّة ونباتات.

Ex. 38

ST: Enclosed a catalogue of our mechanical toys. We will allow a trade discount of 20% and a special discount of 7%.

(From a commercial text)

HT: مرفق فهرس لألعابنا الميكانيكية. وسننصح بخصم تجاري مقداره 20% وخصم خاص مقداره 7%.

NA: مرفق فهرس الألعاب الميكانيكية. سننصح بخصم تجاري من 20% وخصم خاص من 7%.
Ex. 39

ST: The centromere regions of the homologous chromosomes do not pair.

(From a medical text)

HT: 

وكذا منطقة الجزء الرئيسي للكروموسومات المتماثلة لا تزدوج.

NA: 

منطقة الجزء الرئيسي للكروموسومات المتماثلة لا تزدوج.

Ex. 40

ST: The outlays for interest expense are usually shown in this section of the cash budget.

(From a financial text)

HT: 

كما إن المدفوعات في مصروف الفوائد تظهر عادةً في هذا الجزء من الميزانية النقدية.

NA: 

المدفوعات لمصروف فائدة يعرض عادةً في هذا القسم لميزانية النقد.

5.2.2 Linguistic Problems

The linguistic factors are generally most significant and immediate clues as to the fact that a system produces the text. Actually, the complexity of the linguistic structure of the Arabic language seems to have posed somewhat of a handicap for MT systems. The result is often unreadable and intangible translation both from English into Arabic and vice versa. The most typical errors which occur in English into Arabic and Arabic into English translation tend to be structural and lexical, as well as, several semantic and
morphological errors are evident in an MT analysis. In this respect, the texts of the data have revealed a number of linguistic problems peculiar to the grammatical structure of words and sentences and domain-specific terminology. It is worth noticing that these phenomena constitute further causes for influency and inaccuracy of MT output. The main translation problems of this kind that have been encountered in the present evaluation involve:

5.2.2.1 Inadequate Terminology Interpretation

Lexicalization is concerned with choosing the right word or expression. This can go wrong in a number of ways. The most obvious and trivial case is where words are incorrectly and inappropriately associated with underlying concepts. MT systems that suffer such a fault may produce completely nonsensical output. Indeed, any MT system is determined above all by the quality and range of its dictionary information in other words, no MT system will produce good translation if its dictionaries are inadequate.

Despite the low performance of the three Arabic MT systems in the terminology test, the results show that in some cases MA is more successful than NA. But, both are clearly better than GW system for they have the potentiality of making domain-specific/user dictionaries which contribute to extremely better translation quality. It should be confirmed, here, that because the guidelines of this test were precise, raters have been strict in their implementation of them and only restricted themselves to exact comparison with HT. This might be one of the reasons of low systems’ performance in this test. However, various types of problems have been identified in the data of these systems as major causes of inadequate terminology rendition. These are presented in the following points:
(1) **Untranslated Terminology**

Untranslated words are generally a great source of semantic errors and wrong syntactic categorization. This eventually results in unintelligible and unfaithful translation. Even though they use domain-specific dictionaries, many instances of untranslated words (i.e., either missing or left in their English forms) appear in the MT output of both MA and NA (See Ex. 19, 20, 29, 30 and 31 above). On the other hand, this phenomenon is more apparent in the translations produced by GW system as it does not have the option of making user dictionaries. So, this can be considered one of the reasons of poor MT quality of the system (See examples 1, 2, 18, 28 and 29 in the previous sections). It should be stressed, here, that it is evident in the data of MA that this system translates a given term via the user dictionary, but when it occurs in other sentences in the same text, it is left untranslated, i.e., it reappears in its SL form.

(2) **Forms Irrelevant to the TL**

Some ST terminologies occur as deformed words which are irrelevant to the TL in the MT output of the Arabic systems. Especially when they are key words and expressions, they cause alteration and distortion of meaning. This is mainly observed in the data of MA and NA systems as a result of duplicated letters or additional letters to these words. Besides, sometimes strange words appear due to literal rendition. The underlined words in the following extracts manifest this:

Ex. 41

ST: …, increase their lethality … .

(From a military text)

HT: … وتنزّد في فكّها …
Ex. 42

ST: … and with my joint and mutual obligation for all liabilities ….

(From a legal text)

Ex. 43

ST: … we were impressed by your manufactures of cameras ….
   Enclosed a catalogue of our mechanical toys.
   For orders received ….
   … to welcome you as one of our customers.

(From a commercial text)
Ex. 44

ST: … or temporary investments made.

(From a financial text)

HT:

... أو أجراء بعض الاستثمارات المؤقتة.

NA:

... أو استثمارات مؤقتة المجمع.

Ex. 45

ST: … in the wetting and non-wetting fluids.

(From a petroleum text)

HT:

... في الموائع المبللة وغير المبللة.

NA:

... في التبليل وغير – الذبيطل، موائعاً.

Ex. 46

ST: … its imagery lacks the definition to find and characterize some types of targets: …

(From a military text)

HT:

... إذ تفتقد عواكسها إمكانية التحديد لإيجاد وتمييز بعض أنواع الأهداف: ...

NA:

... يتربو صورته عاكسه التحديد من الاكتشاف ويدبت بعض الأنواع لأهدافه: ...
Ex. 47

ST: … deprives the child of skills – enhancing education….

… aims to correct the distorted incentive structure ….

(From a social science text)

HT:

… كما إنه يحرم الطفل من التعليم الذي ينمي مهاراته، …

… أن تهدف السياسات إلى تصحيح بنية الحوافز المنشوهة …

NA:

… يحرم أيضاً الطفل للمهارات – التدريبي تعليماً ، …

… على سياسات أن تتهدف أن تتوجيه المنشوهة بناء محفزاً …

(3) Synonym Usage

One of the major reasons of inexact rendition of words and terminologies in the data of the Arabic MT systems is the heavy reliance on synonyms (i.e., not optimally used words and terms). This is a marked feature especially in the output of GW due to the effect of the user dictionaries with the other systems. In fact, this phenomenon has negatively affected the evaluation results of the terminology test. Some illustrative examples are given below:

Ex. 48

ST: … my moveable and immovable property ….

(From a legal text)

HT:

… أموالي المنقولة وغير المنقولة …
Ex. 49
ST: … of the homologous chromosomes do not pair.

(From a medical text)

Ex. 50
ST: Typical fluid distributions …

(From a petroleum text)
Ex. 51

ST: ..., overflow structures at suitable locations...

(From a water engineering text)

Ex. 52

ST: The central role of social cohesion ....

(From a social science text)
GW:
الدور الـمركزي للتماسك الاجتماعي ....

MA:
رئيسي دور من تلائم اجتماعي ....

NA:
دور التلائم الاجتماعي الرئيسي ....

Notice the effect of the user dictionary on the exact rendition of domain-specific terms regarding MA and NA.

(4) Inappropriate Word Choice
As meaning is heavily dependent on context, contextual information and a broad coverage of word senses are essentially to be encoded in a system’s dictionary. As the data of the three Arabic MT systems reveal, they are unable to use context to increase accuracy of word choice. The MT system must use more information from the context to determine what translation is best. This is a fundamental problem which programmers of these systems must overcome in order to improve the quality of their products. The following excerpts are typical examples (See also, Ex. 28, 29 and 30 above):

Ex. 53
ST: … and the two members of each pair become oriented on the spindle (Fig.1- 2D).

(From a medical text)

HT: ... ويصبح عضوا كل زوج بوضع متكيف على محور الدوران (شکل ۱-۲ د).
GW:

... والاثنان من أعضاء كل زوج يصبح موجهًا على المغزل (تبنيه 1-2 دي).

Ex. 54

ST: As it should be noted in Fig. 4A that the primary focal point $f$ for a converging lens lies to the left of the lens, whereas for a diverging lens it lies to the right.

(From a physical text)

HT:

كما ينبغي أن نلاحظ في الشكل (4 ا) أن النقطة البؤرية الأولية $f$ للعدسة اللامة تقع على يسار العدسة بينما تقع على اليمين في العدسة المفرقة.

MA:

يجب ملاحظة في التنبه 40 الذي أولية نقطة بؤرية $f$ للكانين عدسة لامة على يسار العدسة، بينما لعدسة مفرقة تستلقي على اليمين.

Ex. 55

ST: ... depends on credit terms ...

(From a financial text)

HT:

... تعتمد على شروط الائتمان ....

NA:

... يعتمد عبارات بالدين ....

(5) Mistranslations

It is noticeable in the data of the Arabic systems considered that some bad translations of critical words and terms are attributed to a word-for-word interpretation and transliteration. The result is mostly inappropriate word usage such as, for instance, translating the SL word into two or three words.
This spoils the exact rendition of words and the clarity of meaning. The examples below, illustrate this point:

Ex. 56
ST: I was born in Cairo in 1942 …
(From a legal text)

HT:
ولدت عام 1942 بالقاهرة …

GW:
أنا كنت ولد في القاهرة في 1942 …

Ex. 57
ST: … and drawn up this bond and it has been read aloud to him in the Council.
(From a legal text)

HT:
… ونظمت هذا السند وتم عليه في المجلس علناً.

GW:
… ورسم هذا يونيد وهو قرارا جمهورياً إليه في المجلس.

MA:
… ونظمت هذا يونيد وهو قرارا علنا إليه في المجلس.

Ex. 58
ST: … and the keyboard-TV portion of desk-top computers ….
(From a computational text)
Ex. 59

ST: Stereoisomers are isomers that are identical …

(From a chemical text)

Ex. 60

ST: …, known to me or whose identity was attested to me by Mr. … .

(From a legal text)

Further, some other problems of this kind are found in the data of GW merely, for the positive effect of the user dictionaries with the other systems. These concern the translation of two-word terminologies where one is not translated (i.e., left in its SL form) and the other is badly translated for one reason or other. Also, a kind of instability in the way of translating proper nouns is noticeable here. The following are typical examples:
ST: …the image acquisition module is often referred to as frame grabber.

(From a computational text)

HT: ... كثيراً ما يشار إلى وحدة تحليل الصور بقبابض الإطار.

GW: وحدة استعمال الصورة في أغلب الأحيان مدعومة باسم إطارgrabber

Ex. 62
ST: …the two focal lengths have the ratio of the two refractive indices (See Eq. 3a)

(From a physical text)

HT: ... إذ يكون للبعدين البؤريين نسبة معامل ألكسلي ألكساري (نظر المعادلة 3آ).

GW: (Eqs. 3) refractive ... حيث الطولين البؤريين ليهما نسبة الإثنان فهارس (تريان 3).

Ex. 63
ST: …on the strengths and weakness of the AWACS, JSTARS, U-2, Rivet Joint, P-3, satellite, and other sensors platforms …

(From a military text)

HT: ... عن نقاط القوة والضعف في طائرات البوكس والجستار، والبوتو، والرايفت جوينت، والبيثري، والأقمار الصناعية، ومنصات التحسس الأخرى...
Notice the different renditions given to each of the proper nouns in the last extract reflecting a combination of transliteration and bad literal translation of ‘Rivet Joint’ into ‘فصل مسار’ instead of the ra'iyat joint.

6. Wrong Ordering of Phrasal Constituents of Terms

There are other lexical errors in the Arabic MT sentences of the three systems. Although most of the necessary domain-specific terminologies have been inserted into the user dictionaries of both MA and NA, their translations, in most cases, have not been precise or correct, due to word-order and other grammatical problems. In other words, without such problems, accuracy of terminology rendition of these systems could have been much better.

To avoid any kind of overlapping with the next section which is devoted to detailed identification and discussion of the syntactic problems in the Arabic MT output, no thorough argument in this respect is presented here. These are only dealt with on the level of words and terminologies. Thus, some typical examples showing word-order problems causing inexact and distorted interpretation of terms are given below:
ST: Obligate symbiosis between cellulose-digesting microorganisms and animals may be illustrated by two examples.

(From a biological text)

HT: يمكن إيضاح التكافل المسر بين الكائنات ألمجهرية الهامة للسيلولوز وبين الحيوانات بواسطة مثالين.

GW: المنزيم بين سيلولوزية – هضم كائنات حية مجهرية وحيوانات قد يصوران بمثالين. 

Ex. 65

ST: Similarly, as in a mitotic division, the female as well as the male primitive germ cells (primary oocyte and primary spermatocyte) replicate their DNA just before the first meiotic division begins.

(From a medical text)

HT: وعلى غرار الانقسام الخلوي تضاعف الخلايا الجرمومية الأنثوية والذكورية الابتدائية (وهي خليه البيضة الأولية وخلية البويضة الأولية) محتوى من DNA قبل بدء الانقسام الاحتزالي الأول.

MA: بنفس الطريقة، كما في انقسام الخلوي، الأنثى بالإضافة إلى الذكر الابتدائي (أولى خلايا البيضة الأولية) تضاعف دي إن أي هم مباشرة قبل أول الاحتزالي الانقسام تبدأ.

Ex. 66

ST: The function of the display module is to read an image memory, convert the stored digital information into an analog video signal, ....

(From a computational text)
7. Errors of Category and Word Class

Other grammatical phenomena affecting the precise and appropriate interpretation of words and terminologies involve errors of categories of nouns and verbs (e.g., number, gender, tense, etc) and wrong part of speech (e.g., noun, verb, adjective, pronoun, preposition, etc.). Some extracts from the data of the Arabic MT systems illustrating this point are given below:

Ex. 67

ST: In subsequent stages the members migrate to the opposite poles of the cell (Fig. 1-2)

(From a medical text)

HT: وفي مراحل تالية، يهاجر الأعضاء إلى القطبين المتصادرين للكلية (شكل 1-2).

GW: وفي المراحل اللاحقة، يهاجر الأعضاء إلى الأقطاب المعاكسين للكلية (تينه. 1-2).

It is apparent that there is an error of number in the sentence translated by GW system. The combination 'القطبين المتصادرين' indicating plural should be replaced by the dual form in Arabic 'القطبين المتصادرين.'
Ex. 68
ST: … and holding Jordanian passport No. … issued in …. 

(From a legal text)

HT: 

وحامِل جواز السفر الأردني المرقم ... والصدر في ...

GW: 

وحة ص جواز سفر أردني لا ... أصدر في ...

Here, there is an error of word class in the MT construction. The verb ‘صدر’ should be an adjective ‘الصدر’.

Ex. 69
ST: … by their endosymbiotic association with cellulose-degrading microorganisms. 

(From a biological text)

HT: 

... عبر مراقبته التكافلية الخارجة الأكثر تعقيداً مع الفطر المحطم للسيلوز.

NA: 

... خلال مشاركته التكافلتهم الأكثر تعقيدا الخارجي به فطر محظنة سيلوزي.

There is an error of gender in the MT sentence. The feminine adjective ‘محظنة’ should be replaced by the masculine ‘المحطم’ to achieve appropriate agreement with the preceding masculine noun ‘الفطر’.

Ex. 70
ST: Direct labor and other wage and salary outlays …
An error of category ‘number’ occurs, here. The singular noun ‘راتب’ should be in the plural form ‘رواتب’ to appropriately agree with the preceding noun ‘مصروفات’ or ‘مصرفات’.

Ex. 71
ST: … the same sewer is used for both storm-water and dry-weather flow

(From a water engineering text)

In addition to word order and word choice problems, there is an error of word class, in the MT sentence. The verbal phrase ‘يفجف – تدفق الطقس’ should be altered into a noun phrase ‘المناخ الجاف’, as seen in the HT.

Further, there are other grammatical problems causing deformed terminology such as incorrect determiner usage (e.g., ‘ميزانية النقدية’ instead of ‘الميزانية النقدية’) that will be discussed in the next section. It is worth noticing
that the most problematic phenomena of inexact and unclear terminology rendition in the data of MA are these of word-order and errors of category (as the words are adequately translated by the domain-specific/user dictionaries). Although the something occurs with NA system, it suffers from another problem that is so apparent in its MT output. This concerns the repetition of words and duplication of letters within terminologies. As there are no user dictionaries with GW, the system suffers from more serious problems including untranslated terms, overuse of synonyms and bad rendition of terms.

5.2.2.2 Incorrect Rendition of SL Grammatical Structures

Due to the large differences in the grammatical structures of the two languages, MT between English and Arabic is not an easy task at all. As stated earlier in this chapter, the main source of deficiency is syntax. There are major syntactic errors, in fact, that significantly alter structure and meaning and badly affect the fidelity of the MT output to the original SL text. These errors can originate from several phenomena some of which have been referred to in the previous sections. As a matter of consequence, they, generally, lead to low MT quality.

A detailed analysis of the evaluation results has revealed that none of the Arabic MT systems involved in this study is good at rendering English syntax correctly. It seems clear that more complex restructuring may well be beyond the capacity of the Arabic systems such as, for instance, the treatment of the plural. The English-into-Arabic MT system must be capable of recognizing the equivalence of the structure in Arabic, i.e., when generating the Arabic structure, it must know that Arabic has three ranges for number: single, dual and plural and that English plural must be mapped
into dual in Arabic sometimes. Thus, it is essential that Arabic MT output complies with Arabic grammatical rules and at the sometime preserves its fidelity to the SL structure and meaning. In most cases, this requires the application of four transformational syntactic processes of rearrangement, addition, deletion and replacement.

A further and more detailed identification and discussion of Arabic MT syntactic problems is presented in the following points:

1. **Bizarre Syntactic Arrangement**

In spite of the serious lexical problems of untranslated words and systems’ inability to handle polysemy, the major system’s handicap in Arabic MT (causing confused structure, poor style and awkward translation) is that of word order. The free word order nature of Arabic makes parsing of sentences so difficult. In fact, the three participating Arabic systems fail to produce sentences that exhibit the TL syntactic rules and arrangements. In most cases, sentences start with the noun and place the verb towards the very end position preserving the SL word order which is the reverse order of the TL. This problem has been observed at all grammatical levels of the Arabic MT sentences in the data (i.e., clausal and phrasal). Even when words and terms are appropriately selected on the basis of context, this problem violates the form and meaning of the whole sentence. Thus, sound post-editing processes concerning rearrangement of syntactic structures of sentences are badly needed to improve the quality of Arabic MT output. In addition to the extracts given in the previous sections, the one below is a typical example:
Ex. 72

ST: The idea of a single platform to perform the functions of the AWACS and JSTARS is also being explored.

(From a military text)

HT: وتدرس حالياً فكرة وجود منصة واحدة لإنجاز مهام الأواكس والجستار.

GW: فكرة رصيف وحيد لإداء وظائف أواكس وJstars أيضاً أن يستكشفا.

MA: فكرة منصة واحدة لإداء اعمال الأواكس وجستارز أيضاً ان يدرس.

NA: فكرة منصة فريدته أن يقوم بمهام أي دبليو أي سي إس وجستارز أيضاً تدرس.

2. Lack of Grammatical Agreement

Another phenomenon causing incorrect rendition of MT sentences concerns syntactic agreement or concord between the sentence elements. All kinds of differences in gender, number or case are another source of relatively frequent errors. Arabic is a language with strong requirements of gender, number and case agreement not only between subject and verb (gender and number agreement), but also in several other types of constructions. In the data of the present evaluation, it is found that no such agreement exists between the sentence elements in most cases. The three systems reveal a considerable failure in this respect, as the following extracts show:

Ex. 73

ST: The hydrocarbon fluids which occur in a given trap depend on ….

(From a petroleum text)
There is a lack of gender agreement between subject and verb, here, as the verb 'should be' to match the feminine noun functioning as subject.

Ex. 74

ST: … and separates into two daughter chromosomes which migrate to opposite poles of the cell (Fig. 1-1 D, E).

(From a medical text)

There is a lack of number agreement among noun, adjective and verb in the translation of MA for the wrong rendition of ‘chromosomes’ and ‘migrate’ into ‘يهاجر’ without taking the dual form into consideration. Also, there is a similar case in the translation of NA indicating lack of number agreement between noun and adjective where the latter is translated into ‘بنوين’ a singular form instead of the dual ‘بنوين’.
Such a problem stems from the syntactic discrepancies between two languages when rendering this grammatical aspect.

Ex. 75
ST: … provided that there are sufficiently strong institutions … social groups.

(From a social science text)

HT :

MA:

Here, the lack gender agreement between nouns and adjectives is due to word-for-word translation without taking into account the syntactic rules of the TL. English, in fact, has three genders: masculine and feminine (for animate nouns) and neuter (for inanimate nouns); while Arabic has only two: masculine and feminine for animate and inanimate things. Moreover, very few nouns are marked for gender in English. In Arabic, on the other hand, most nouns are marked for gender which is also relevant to pronouns, verbs and adjectives.

Ex. 76
ST: …, these fungi are able to thrive on fresh leaves …

(From a biological text)

HT :

... فإن هذه الفطريات تكون قادرة على النمو على أوراق طريه ...
The MT construction of GW shows incorrect number rendition where the singular noun ‘فطر’ should have been translated into plural ‘قطريات’. Ex.77

ST: There is no need for overflow structures, which are relatively costly.

(From a water engineering text)

HT:
ولا يوجد داع لمنشآت الطفح التي تكون باهظة الثمن نسبياً.

MA:
ليس هناك حاجة لمنشآت الطفح الذي نسبياً غالي.

The sentence translated by MA, here, reveals incorrect gender rendition on the level of relative pronouns which is usually determined by the gender of the antecedent.

Ex. 78

ST: Thus this module is often called ....

(From a computational text)

HT:
ولنها غالبًا ما تسمى هذه الوحدة ...

NA:
هكذا هذه الوحدات غالبًا ...
There is an error of gender for the inappropriate choice of the singular demonstrative pronoun ‘هذا’ instead of ‘هذته’ by NA system. This is due to word-for-word translation irrespective of context and TL (here, Arabic) syntactic rules, for Arabic demonstratives have gender distinction in singular and dual-the thing not found in English.

Ex.79

ST: … the homologous pair consists of four chromatids (Fig. 1-2 B).

(From a medical text)

HT:

... فإن الزوجين المتماثلين يتآلفان من أربع كروماتيدات (الشكل 1-2 ب).

NA:

... يتآلف الزوج المتماثلة من اربع كروماتيد (شكل 1-2 ب).

Notice how NA system is incapable of correct number rendition as it has translated the dual form ‘الزوج المتماثل’ into singular ‘الزوجين المتماثلين’ (here, there is also lack of gender agreement between noun and adjective); and the incorrect rendition of the singular noun ‘كروماتيد’ instead of the plural ‘كروماتيدات’.

3. Differences in Cases
In general, Arabic has three cases (nominative, accusative and genitive) which are expressed explicitly in nouns, pronouns and adjectives. English, on the other hand, has two cases for nouns (common and genitive), three for pronouns (nominative, accusative and genitive) and no case for adjectives (Aziz, 1989:132). These cases are distinguished by diacritics, morphological inflections and prepositional constructions in Arabic. For example, the Arabic equivalents for ‘an engineer’ and ‘engineers’ (the singular and plural forms for all cases in English) are: nominative مهندسَن, مهندس and accusative/ genitive مهندسَن, مهندسَن. Such differences in cases, of course, result in translation errors, because the main bilingual dictionaries of the Arabic MT systems do not contain valency and subcategorization information. The following are typical extracts culled from the data of the three participating systems:

Ex. 80

ST: We have an experience of more than thirty years ....

(From a commercial text)

HT:

لدينا خبرة تزيد على ثلاثين عاماً ....

GW:

عندنا تجربة أكثر من ثلاثين سنة ....

MA:

عندنا خبرة أكثر من ثلاثين سنة ....
In all the above MT sentences, the genitive case should have been used instead of the nominative.

4. Incorrect Pronoun Translation
This is another known area of deficiency for Arabic MT systems. Again, the data of the three participating systems manifest certain incorrect pronoun translations mainly due to differences in gender and number, homographs and other problems in addition to the complexity of the Arabic pronouns system. The English personal pronoun ‘you’ for instance, is rendered into Arabic in no less than five ways: انت / انت / أنتما / أنت / أنتت. Below are some examples of this case:

Ex. 81
ST: The primary focal point F is an axial point having the property that any ray coming from it, or proceeding toward it, travels parallel to the axis after refraction.

(From a physical text)

HT: و تكون النقطة البؤرية الرئيسة (F) نقطة محورية تنصف بأن أي شعاع يصدر منها أو يمضى نحوها ينتقل موازياً للمحور بعد الانكسار.

GW: إن النقطة المركزية الأساسية إف نقطة محورية سيكون عدتها الملكية التي أي مجيء شعاع منه ، أو
For gender disagreement between the sentence elements, pronouns in the above MT sentences whether separable or inseparable are incorrectly translated as the masculine form has been used instead of the feminine. The next extract shows this problem as a result of incorrect number rendition where the plural forms of the possessive pronoun ‘your’ and the personal pronoun ‘you’ should have been selected in the Arabic MT sentences.

Ex. 82

ST: … and were impressed by your manufactures of cameras….

I send you a letter ….

(From a commercial text)

HT: ... وأعجبنا بالآلات التصوير التي نصنعها ... 

أرسلت لكم رسالة ....

GW: ... وكانت معجبه بصناعاتكم من آلات التصوير ....
In most cases, the relative pronoun “that” is incorrectly rendered into the demonstrative pronouns ‘ذلك’ or ‘تلك’. Also, it is sometimes wrongly translated into the singular forms ‘التي’ or ‘الذي’ where the dual or plural is required, as the following examples illustrate:

Ex. 83
ST: … for the level of progress that can be achieved in “closing the loop” in the future.

(From a military text)

HT: … مستوى التقدم الذي يمكن إحراره في مجال ’أهداف الحلقة ‘ في المستقبل.

NA: … مستوى التقدم تلك يمكن أن ينجح في ’طوق ’ الحلقة ‘ في المستقبل .

Ex. 84
ST: Two compounds that constitute a pair of cis-trans isomers may have different common names, e.g., maleate and fumarate.
(From a chemical text)

HT: وقد يكون للمركبين اللذين يكونان زوجين من الأشباه الجزيئية المقربة -المفرقة اسمان شائعان مختلفان مثل المالات و الفيورمات.

GW: مركبان التي تشكلان زوج isomers عبر transcis لربما له أسماء مشتركة مختلفة، و مثال على ذلك: maleate و fumarate.

MA: مركبان التي تكونان زوجي من الأشباه الجزيئية المقربة -المفرقة لربما عنده أسماء مشتركة مختلفة و مثال على ذلك: - مالات و الفيومولات.

Yet, another cause of incorrect pronoun rendition concerns the inseparable explicit pronoun ‘يَتَّ’ attached to the past verb form in Arabic with the function of subject (e.g., it stands for the implicit pronoun ‘أنا’, the subject as in ‘درست’). As English pronouns are only explicit and separable, a sentence like ‘I received my BA in Humanities …’ is literally translated by MA system like ‘أنا حضر ليسانس في العلوم الإنسانية …’ irrespective of the syntactic rules of Arabic. This indicates a deficiency in the syntactic component and transfer rules of the Arabic MT systems involved. The example below illustrates this point:

Ex. 85
SL: I received my BA in Humanities …
For unknown reasons, however, these systems show a kind of inconsistency in this respect such as the case with GW system in its correct rendition of the verb in the above sentence ‘استللمتُ بي أي ي في العلوم الإنسانيه’ . Nevertheless, the same system has incorrectly translated the next sentence as ‘أنا أستخدمت’ instead of ‘عملت في وظائف متعددة’ for the English sentence ‘I was employed in a number of jobs …’.

5. Wrong Part of Speech Categorization
An obvious syntactic error in the data of the three Arabic MT systems that results in mistranslation and alteration of structure and meaning of the SL sentence is when the wrong part of speech is selected in the TL output. Examine the following representative excerpt:

Ex. 86
ST: As in the case of nitrogen-fixing bacteria and legumes ….
Here, because of word-for-word translation and wrong interpretation of ‘fixing’ into the verb ‘ثبت’ instead of the adjective ‘مثبتة’, the whole meaning of the construction is altered where the reverse is true.

6. Inconsistent Preposition Handling

The problem of Arabic MT systems’ inability to handle prepositions correctly and appropriately is quite serious. Often, the wrong preposition appears in the MT sentence, two successive prepositions are used, there is a missing preposition or there is a need to delete one. The following are typical examples:

Ex. 87

ST: … and have a good knowledge of the market here …

(From a commercial text)
GW: ... وله معرفة جيدة من السوق هنا ...

In the above MT construction translated by GW system, the replacement of by ‘بِ’ is needed.

Ex. 88
ST: The gravitational forces cause the less dense fluids to seek the higher positions in the trap.

(From a petroleum text)

HT: إذ تدفع قوى الجاذبية الموائع الأقل كثافة إلى الصعود إلى المناطق العليا من المصيدة.

MA: القوى الجاذبية تدفع الموائع القل كثافة إلى الصعود، المناطق العليا في المصيدة.

A look at the HT apparently shows that the preposition ‘إلى’ is needed to precede the noun phrase ‘المناطق العليا’.

Ex. 89
ST: I received my BA in Humanities from the Department of Islamic Oriental languages University, in 1965 with the grade of ‘Good’.

(From a commercial text)

HT: وفي عام 1965 حصلت على بكالوريوس الآداب في الإنسانيات من قسم اللغات الشرقية الإسلامية بدرجة ‘جيد’.
The deletion of ‘لـ’ before ‘جيد’ is needed in the above MT sentence.

It is worth noticing that in the translation of MA system in Ex.90 below, there are two successive preposition ‘في’ and ‘بـ’ (which is inseparable). By comparison with the HT, the first should be omitted.

Ex.90
ST… and the compound can be drawn in 2 ways that are mirror images of each other.

(From a chemical text)

HT: ...و يمكن رسم المركب بطريقة تكون كل واحدة منهما صورة متطابقة للأخرى.

MA: ...و المركب يمكن أن يرسم في بطرقتين ذلك صورتان مراويتان كل منهم الآخر.

7. Failure to Treat the Definite Article

The structure of the noun phrase in Arabic, like that of English, is one of modification. In the basic noun phrase, modifiers are mainly determiners (e.g., articles, possessive and demonstrative pronouns). The definite article ‘الما’ is normally used with a noun to indicate generic reference (e.g.,...
‘water’. In other cases, the definite article ‘ال’ is also used to express specific or unique reference (Aziz, 1998: 101-106). Errors regarding the rendition of the Arabic definite article are evident in the MT output of the three evaluated systems. This is an indication of the systems’ incapability to treat this syntactic aspect appropriately due to the missing definite article in the Arabic text or vice versa, i.e., it is arbitrarily added by the system where it does not fit. Moreover, it is even interpreted wrongly in some instances where it is replaced by the preposition ‘ل’. Examine the following representative examples:

Ex. 91
ST: The influence of a combined system on a wastewater treatment plant and the environment must be considered when selecting a sewer system. (From a water engineering text)

HT: كما يجب أن يؤخذ بنظر الاعتبار تأثير المنظومة المشتركة على محطة معالجة مياه الفضلات وعلى البيئة عند اختيار منظومة المجاري.

GW: تأثير نظام مشترك على نبات معالجة مياه فضلات البيئة يجب أن يؤخذ بنظر الاعتبار عندما يختاران نظام بالوعة.

Since there are only two articles in Arabic, the definite article ‘ال’ and zero article (often plus nunation التنوين), the Arabic MT systems are unable to recognize the English indefinite article ‘a’ as it does not have an equivalent in Arabic. They substitute it by the zero article in the Arabic
texts. As a consequence, it would be missing in the translation as the above example shows when rendering ‘a combined system’ into ‘نظام مشترك’, for instance, instead of ‘المنظمة المشترك’.

Ex. 92
ST: Prior to the disturbance of the accumulation….

(From a petroleum text)

HT: وقبل حدوث الاضطراب في التجمعات النفطية…

MA: قبل اضطراب للتراكم …

The phenomenon of replacing the definite article ‘ال’ by the preposition ‘ل’ as in the Arabic MT output of MA system in the above example (where ‘التراكم’ should be replaced by ‘التراكم’) often occurs when the noun phrase functions as the complement of preposition in the SL sentence.

Ex. 93
ST: Cash receipts depend on collections of accounts receivable ….

(From a financial text)

HT: وتتعد المقابلات النقدية على تحصيل حسابات المدنيين …
Here, NA system has incorrectly attached the definite article ‘الـ’ to the nouns ‘تحمل’ and ‘حسابات’ where it does not fit and should be deleted.

8. Problems of Tense, Aspect and Voice
In general, the same verb forms used to express tense are also used to express aspect in Arabic; while aspects in English are formed by combinations of verbal elements (i.e., a form of auxiliary ‘have’ plus the past participle form of the lexical verb). Moreover, the passive verb in Arabic is basically a simple phrase like that of the active sentence and the transformation from active to passive is morphological rather than syntactic (e.g., كتب / كُتِبَ; wrote / was written). In English, the verb phrase in the passive belongs to a complex structure comprising the auxiliary ‘be’ and the past participle of the lexical verb. For such differences between the two languages, some errors regarding the grammatical categories of verbs occur in the data of the Arabic MT systems mainly attributed to certain deficiencies in their syntactic information and transfer rules.

For both the passive and perfective tenses, the Arabic MT systems often fail to treat the English verbal combination as one unit. Rather, they translate each verb separately (i.e., the auxiliary and the lexical verb) into its equivalent in Arabic as a matter of word-for-word rendition. Some typical extracts are given below:

Ex. 94
ST: … and were impressed by your manufactures …

(From a commercial text)

HT: ... وأعجبنا بصناعاتكم ...

GW: ... وكانت معجبة بصناعاتك ...

Ex. 95
ST: … and acknowledges its contents and entirely confirms its correctness and has signed it ….

(From a legal text)

HT: ... وااعترف بمضمونه ووافق على صحته تماماً ووقعه ...

MA: ... واافتر واعترف مضمونه وكلياً وافق علاه صحت وعندى وقعه هو ...

Ex. 96
ST: I was employed in a number of jobs in Cairo ….

(From a legal text)

HT: عملت بوظائف متعددة بالقاهرة ...

MA: أنا كنت عملت في متعددة من وظائف في القاهرة ...
ST: I have therefore convened a Notary Council … 

(From a legal text)

HT: 

وعليه عقدت مجلساً للعدل …

NA: 

عندى لذلك أقـيل مجلس للعدل …

Besides, these MT systems occasionally fail to appropriately handle the passive construction according to the grammatical rules of the TL, as illustrated in the examples below:

Ex. 98
ST: Usually, several modules are incorporated in an image processor.

(From a computational text)

HT:

وتُدمِج عادةً وحدات كثيرة في معالج الصورة.

GW:

 عدة وحدات متحدة في معالج صورة

MA:

 عادة، عدة وحدات متحدة في معالج صوره

Ex. 99
ST: a. The two systems must be compared ….

b. Those sewers which carry sewage (dry-weather flow) are called sanitary sewers.

(From a water engineering text)
In the data of the Arabic MT systems, it is also noticeable that such systems occasionally fail to translate the correct tense into the TL. This phenomenon mainly occurs with more complex tenses resulting in the wrong Arabic tense from the syntactic and semantic viewpoint. The following are some representative examples:

Ex. 100
ST: I sent you a letter on 18 August asking for a hundred dozen of fountain pens model … that would be delivered within two weeks.

(From a commercial text)

HT: أرسلت لكم رسالة في 18 آب طالباً فيها من ذيته أقفام حبر نموذج... التي كانت ستسلم خلال أسبوعين.

GW: أرسلت رسالته في 18 أغسطس / آب يسأل عن ذيته ومانة من نموذج أقفام الحبر... الذي سيسلم خلال أسبوعين.

Notice how the incorrect tense rendition of ‘would be delivered’ into ‘سيسلم’ ‘كانت ستسلم’ changes the time of the action from past into future.
Ex. 101
ST: Total cash needed includes ....

(From a financial text)

HT:

ويشمل إجمالي النقد المطلوب ....

NA:

نقد كلي مطلوب شمل ....

Here, the present tense in the ST is incorrectly rendered into past in the MT output of NA system.

Further, despite the fact that the copula is normally not expressed in Arabic (unless it is required to indicate the past or the future) and it is possible to find complete sentences in this language without a verb element, it is necessary sometimes to render this verb into a form of ‘يكون’ in the TL for stylistic purposes. See how the following MT sentences would have sounded better if verb ‘be’ had been translated in the output sentence:

Ex. 102
ST: Social cohesion can be high when a society is relatively homogenous ....

(From a social science text)

HT:

ومن الممكن أن يزداد التلاحم الاجتماعي عندما يكون المجتمع متجانساً نسبياً ....

GW:

التماسك الاجتماعي يمكن أن يكون عالي عندما مجتمع متجانس نسبياً....

Ex. 103
ST: If they are on opposite sides of the double bond, the compound is the trans isomer.

(From a chemical text)

Ex. 104
ST: The axis in each case is a straight line …

(From a physical text)

In sum, there might be other syntactic problems, but the ones discussed above are the most obvious and the major causes of inefficient and inadequate MT output. This is mainly attributed to the systems’ failure to deal with the Arabic grammatical rules and certain deficiencies in the translation process itself. The overall effect is that of limited and poor syntactic quality of TL texts.

5.2.2.3 Erroneously Recognized TL Morphology

It is a truism to say that one of the most straightforward operations of any MT system should be the identification and generation of morphological
variants of nouns and verbs. There are basically two types of morphology in question: inflectional morphology, as illustrated in the familiar verb and noun paradigms (e.g., English: ‘write’, ‘writes’, ‘writing’, ‘wrote’ and ‘written’; and Arabic: ‘تكتب’، ‘كتبت’، ‘كتبت’، ‘كتب’، ‘كتب’، ‘كتب’، ‘كتب’، ‘كتب’، ‘كتب’، ‘كتب’، ‘كتب’، etc), and derivational morphology, which is concerned with the formation of nouns from verb bases, verbs from noun forms, adjectives from nouns, and so forth (e.g., English: nation, nationalism, nationalize, nationalization and their equivalents in Arabic: تأميم يومم قومية قوم).

It is worth stressing that any MT system should as a minimum be capable of recognizing morphological forms and of generating them correctly. In general, an MT system which cannot go beyond morphological analysis will produce little more than word-for-word translations. It may cope well with components and other fixed expressions, it may deal adequately with noun and verb forms in certain cases, but the omission of any treatment of word order will give poor results. Nevertheless, another serious and very complex linguistic problem that results in generating ill-formed MT sentences concerns incorrect TL morphological structure.

An important consideration in the development of Arabic MT systems is the matter of Arabic morphology. Most of the researches in Arabic NLP systems mainly concentrate on the field of morphological analysis (Ditters, 2001 and Othman, Shaalan and Rafea, 2003). In the present work, it is found that all the Arabic MT systems exhibit uniformly low morphological representations (See Figs. 5.5 and 5.6 in 5.1.1.1.2). This is because they are incapable of recognizing the morphological differences and their equivalent structures between the two languages for certain defects in their
morphological analyzers and rules and in the grammatical information contained in the lexicon of each system.

As Arabic is highly derivational (ibid.) and uses to a great extent inflected forms, it is rich in its morphological changes. Most of the basic parts of speech like nouns, verbs and adjectives, change according to tense, aspect, mood, voice, gender, number, case, person and definiteness which express grammatical relations (there is a close relation between the morphological and syntactic structure). Being mainly classified into inflectional and derivational affixes, the morphological structure of Arabic is sub-classified into: prefixes (e.g., 'ن', 'الوطن', 'homeland' and in 'تعيش', 'to live'), infixes (e.g., 'ال', 'كتاب', 'a book' and 'جلّس', 'to sit with') and suffixes (e.g., 'ون', 'يكتبون', 'are writing' and 'مهندّسون', 'engineers'). In English, affixes are of two kinds merely: prefixes (e.g., 're-' in 'reconsider' and 'dis-' in 'disagreeable') and suffixes (e.g., '-s' in 'girls' and '-ize' in 'realize'). Infixes are rare in English and are most commonly replacements not additions. Some replacive allomorphs are called 'infixes' because they are positioned within a word, as in 'spin', 'spun' and 'foot', 'feet' (For more details about English, see Stageberg, 1965: 104).

On the basis of the above classification, some morphological errors have been identified in the data of the three evaluated Arabic MT systems. These mainly result from the lexical and syntactic problems which are discussed in the previous sub-sections in this chapter. In order to avoid delving into further details and complexities of the morphological structure in both
languages (which are out of the scope of the present work), the morphological problems will be discussed only within the framework of this classification. For emphasis, various lexical errors such as: untranslated terms (e.g., ‘isomers’), transliterations (e.g., ‘bond’ into ‘سنده’ instead of ‘سنده’), inability to handle polysemy (e.g., ‘system’ into ‘أنظمة’ instead of ‘أنظمة’), translating into two words instead of one (e.g., ‘two pairs’ into ‘الثنان’ instead of ‘ الزوجي’), wrong word choice (e.g., ‘focal plane’ into ‘طائرة’ instead of ‘سطح بؤرية’), and others have badly affected the correct morphological evaluation of words and terms in this study. However, these will not be discussed here, since the morphological errors within the present classification are not apparent in such phenomena.

Three types of morphological problems have been found recurrent in the data. These are discussed with some illustrative examples below:

1. **Inappropriate Choice of Prefixes**

One major factor that often impedes correct morphological forms in the Arabic MT output is that of definiteness, i.e., the Arabic definite article ‘ال’ which is found absent where it should be used or vice versa. Other problems relevant to the attachment of prefixes to lexical items involves: errors in word class categorization, gender, tense and doubling of prefixes. The following excerpts are typical examples:

Ex. 105

ST: The axis in each case is ….

(From a physical text)
Ex. 106
ST: Conversely, the memory can be addressed …

(From a computational text)

In the two extracts above, the definite article 'ال' has been mistranslated as it needs omission in the first; while it has to be attached to the noun 'ذاكرة' in the second.

Ex. 107
ST: Cash budgets help avoid …

(From a financial text)
The alteration of the part of speech of the verb ‘help’, ‘تَسَاءَد’ in the above MT sentence of GW system causes an error in the morphological structure of this word where the verbal inflectional prefix ‘كِتَ’ of the imperfect verb should have been applied, not to mention the other morphological and syntactic errors.

Ex. 108

ST: Prime examples are mycorrhizae (=fungus – root), comprising the mycelia of fungi ....

(From a biological text)

HT: ومن الأمثلة الرئيسية على ذلك، الفطريات الجذرية (فطر – جذر) التي تشتمل خيوط الفطريات ....

MA: الأمثلة الأساسية الفطريات الجذرية (فطر – جذر)، يشمل خيوط للفطر ....

In this example translated by MA system, there is a lack of gender agreement between the subject ‘الفطريات’ which is a feminine noun and the verb ‘يشمل’ as the wrong choice of the verbal inflectional prefix ‘يَيْ’ (used with masculine noun subject) instead of ‘كِتَ’ indicates.

Ex. 109

ST: Other costs-depend on timing and credit terms.

(From a financial text)
Here, NA system has incorrectly chosen the perfect tense in its Arabic translation of the SL verb ‘depend’, ‘اعتمد’ instead of the imperfect verb form ‘اعتمد’, where the verbal inflectional prefix ‘ات’ should have been replaced by ‘ت’.

A marked morphological error in the translation of NA system concerns the duplication of the prefix attached to a certain lexical items or the use of two different prefixes (one correct and the other incorrect), as seen in the examples below:

Ex. 110
ST: These factors block ….
                  (From a social science text)

HT:  وتنفع هذه العوامل  

NA:  تنفع هذه العوامل  

Ex. 111
ST: … for a converging lens lies …, whereas for a diverging lens it lies …
                  (From a physical text)
For unknown reasons, this MT system even selects the wrong prefix sometimes. The following extract shows this:

ST: The US was able to “close the loop” ....

(From a military text)

HT: وكانت الولايات المتحدة قادرة على ألقاف الحلقة" ....

NA: تمكنت الولايات المتحدة من ان تقفز الحلقة’ ....

2. Errors of Word Infixes

Due to certain syntactic processes involving the derivation of lexical items (e.g., deriving the plural form صواريخ ‘missile’), some morphological changes in the internal structure of words occur. Very few errors of infixes (i.e., the morphological inflections that occupy the medical position of words) have been observed in the data of the participating MT systems. These mainly concern errors in the derivation of plural forms of nouns and incorrect choice of part of speech. Some illustrative examples are given below:

Ex. 112
Many common optical devices contain not only mirrors and prisms ….

(From a physical text)

لا تحتوي الكثير من الأدوات البصرية الشائعة على المرايا والمؤشرات فحسب ….

(From a social science text)

تحتوي العديد من الأدوات البصرية المشتركة ليست فقط شقه المؤشرات والمرايا ….

In these two examples, it seems obvious that the MT systems fail to handle the irregular plural (or the broken plural) in Arabic as this type is not found in the English (SL) plural system. Thus, they try to choose the equivalent form to the SL word which is often treated regularly. As a result, wrong word forms that are irrelevant to the TL appear in the Arabic MT output, such as وَمُؤَشِّرَاتَ ‘ (representing a regular feminine plural) and وَمُؤَشِّرَاتَ ‘ (representing a regular masculine plural). The former should have been translated into 'when the vowel letter 'ا' is to be added to the middle of the word as an infix and the vowel letter of the base form 'و' is to be replaced by 'ي'. In the latter, on the other hand, the vowel letter 'ي'
(here duplicated by the Arabic diacritic ‘شد’ wrongly considered an infix), in addition to the suffix ‘ين’ (indicating accusative case for the regular plural noun) should be omitted.

Ex. 114
ST: Typical fluid distributions resulting from the equilibrium of these forces are shown ….

(From a petroleum text)
HT: إن التوزيع النموذجي للموائع الناتجة عن توازن هذه القوى مبين …. 
MA: التوزيعات أسلاله النموذجية ننتج عن توازن هذه القوى تظهر …. 

Arabic adjectives are mainly derived from verbs and nouns and among the most common adjectives derived from verbs are these of the measure (فاعل) as in the above example ‘ناتج’ which is derived from the root of the verb ‘نتج’ (فعل). In the derivation process, the vowel letter ‘ا’ should be added to occupy the medial position of the word as infix. It seems clear that MA system has not chosen the correct word by treating the SL non-finite verb form ‘resulting’ as an imperfect verb ‘نتج’ instead of an adjective in the Arabic MT output.

3. Inability to Maintain Correct Suffixes
The morphological structure and formation of Arabic words is governed to a great deal by the syntactic relationships among the sentence elements in
terms of agreement in number, gender, case, tense, etc. Suffixes determine the grammatical significance of lexical items which is realized by different morphological inflections and is expressed differently in each language. The noun ‘engineer’, for instance, can be changed into the plural form by attaching the inflectional suffix ‘-s’ to the end, as in ‘engineers’; while in Arabic the plural form of ‘مهندس’ is realized by the inflection ‘-ين’ for regular plural masculine nouns. The following extracts culled from the data of the three Arabic MT systems manifest some suffixation problems:

Ex. 115
ST: … each chromosome undergoes … and separates into two daughter chromosomes ….

(From a biological text)

HT: ... يعاني كل كروموسوم ... حيث ينفصل إلى كروموسومين بنويبين ...

MA: ... بعدنذ كل الكروموسوم و ... separates في كروموسومين بنويبة ...

Here, since the dual plural is not realized in English, MA system has failed to handle it in its translation of the above example. Thus, following the SL grammatical system and rules, the noun phrase ‘two daughter chromosomes’ is translated into ‘كروموسومين بنويبين’ instead of ‘كروموسومين بنويبة’ showing an apparent absence of the suffix ‘-ين’ an inflection attached to Arabic singular nouns to indicate dual plural.
Ex. 116
ST: I, the undersigned …, …

(From a legal text)

HT:

انا الموقع في ادناه …

GW:

انا، الموقعون أدناه …

Here, GW system has rendered the SL noun ‘undersigned’ incorrectly by attaching the inflectional suffix of the Arabic masculine plural ‘و–’ to the end of translated noun in the TL.

Ex. 117
ST: Sewerage systems are of two types: combined and separate.

(From a water engineering text)

HT:

وتكون منظمات المجاري على نوعين: مشتركة ومنفصلة.

GW:

أنظمة تصريف مياه المجاري من نوعين: المشترك والمنفصل.

MA:

منظمات المجاري من نوعين : مشتركة وينفصل.

The above extract illustrates a double error in gender and in part-of-speech categorization. This is reflected in the mistranslation of MA system of the SL adjective ‘separate’ into a verb form ‘ينفصل’ in the Arabic MT output. Consequently, the resultant word lacks gender agreement with the preceding noun ‘منظمات’ that it should have modified (if translated correctly into ‘منفصلة’). This is indicated by the absence of the inflectional suffix ‘ة’.
which is a feminine marker. This also seems clear in the MT output of GW system.

Ex. 118
ST: As women number..., neglecting their capabilities ....

(From a social science text)

HT:
ولما كان عدد النساء ... فإن اهمال قدرائهن ....

NA:
عندما تعد النساء ... الذين يهملون قدرائهن ....

As Arabic possessive system is more explicit than English with respect to the second and third person plural, i.e., it has masculine and feminine gender; the noun phrase ‘their capabilities’ in the SL sentence above could be translated into ‘قدراتهن’ or ‘قدراتهم’. This is, of course, determined by context. Referring to the noun ‘women’ / ‘نساء’, the latter alternative is the correct form. Here, NA system has wrongly rendered this noun phrase by attaching the inseparable possessive pronoun for masculine plural ‘هم’ – ‘هم’ قدراتهن, (which is basically a determiner and realized as suffix) to the noun ‘قدراتهم’, i.e., ‘قدراتهن’ instead of ‘قدراتهم’ (where the suffix ‘هن’ – ‘هن’ stands for feminine plural).

Ex. 119
ST: …, increase their lethality, and reduced both the risk of friendly fire and collateral damage.

(From a military text)

HT: 

NA:

A more common feature with NA system is the use of two successive suffixes of the same type, but each with a different function or the doubling of the same suffix. These phenomena are apparent in the above sentence translated by this system. The first error concerns the word ‘فتكهاهم’ where the second suffix ‘-هم’ is wrongly attached to the noun and should be omitted. The inseparable possessive pronoun for feminine singular form ‘-ها’ is the correct suffix here. As for the second error, one of the two feminine suffixes ‘-ة’ attached to the end of the noun ‘الصديقة’ is to be deleted. Such instances are also noticeable in the translations of MA system (e.g., ‘توظيفهاهم’, ‘رائتنا’, ‘دراسةي’, ‘ملائمته’, ‘مصممونه’, ‘جذورهاهم’).

Further, some suffixation errors are due to choosing the incorrect case and tense for nouns and verbs. The following are typical examples:

Ex.120

ST: Carbon plays a central role in biochemistry ....

(From a chemical text)
As it is always the case in Arabic, adjectives agree with the noun head in number, gender, and case. In the correct translation (HT) of Ex. 121 above, the noun and its adjective ‘دوراً مركزيَا’ are in the accusative (or objective) case which is indicated by the inflectional suffix ‘أ’. Thus, the translation suggested by MA system contains a suffixation error resulting from their failure to realize the correct case (here, the nominative case is indicated by the MT system ‘أ’ instead of ‘أ’). In the next example translated by GW system, i.e., Ex.122, the noun ‘هدفين’ occurs in the genitive case in the HT. This is indicated by the dual number inflectional suffix ‘إين’. Thus, the translation suggested by GW contains a suffixation error resulting from its failure to realize the correct case (here, the accusative on the basis of the translation suggested by this MT system). This is reflected in the
attachment of the wrong suffix 
-ان‘، used with the nominative (or subjective case) as in the noun ‘هدفان‘، which should be ‘هدفين‘.

Ex. 122
ST: It largely failed operational testing before the Afghan conflict with some eight crashes in the six months before the conflict.

(From a military text)

HT: لقد فشلت فشلاً ذريعاً في الاختبارات العملية قبل النزاع الأفغاني إذ تحطمت ثماني مرات تقريباً في الأشهر الستة التي سبقت اندلاع النزاع.

NA: فشلت إلى حد كبير اختبار عملية قبل يتحطم النزاع الأفغاني مع بعض الثمانية ي الأشهر الست قبل النزاع.

As a result of following a word-for-word translation, an apparently poor style is noticeable here. Accordingly, the wrong tense has been selected by NA system.

In this MT output, the imperfect Arabic tense of the verb ‘يتحمل‘ should have been replaced by the prefect verb form ‘تحملت‘ with the inflectional suffix ‘-ت‘ indicating feminization.

In fact, problematic aspects of the Arabic morphological structure are dominant features in the data of the three participating MT systems. These represent one of the major causes of ill-formed words and expressions and poor MT quality, in general.
5.2.3 Operational Problems

In addition to the cognitive and linguistic problems that directly contribute to poor MT quality, other types of deficiencies regarding the operation of the Arabic MT systems and the translation process itself have also been recognized. Two types of such problems have been well identified. The first is related to the measurement of the translation speed; while the second concerns the design of the user dictionaries. These are discussed in detail below.

5.2.3.1 Impediments in Measuring Speech

In measuring the speed of translation some errors hindered the process with NA system merely. This is mainly attributable to the fact that the system currently available in the local market is a copy and not the original version. Therefore, during the process of translation, the system does not work adequately and suffers from constant pauses.

The evaluation indicated that with certain text types, such as the legal, commercial and biological, NA paused and interrupted the translation process with the appearance of a short message confirming that “an illegal operation had been performed.” Immediately after that, the MT system completely shut down and was restarted again. In order to overcome this problem, the part of the text (i.e., a sentence or more sentences) at which the system had paused was deleted. Only then, the system could resume the translation of a given text. In addition to such operational problems of the system itself, other possible causes of this phenomenon concern the existence of long and syntactically and/or semantically complex sentences within these texts. A less serious problem that also impeded the continuity of
the translation process and delayed the system’s performance was related to the system’s pauses for few seconds only to continue the translation thereafter.

Further, in other cases, during the translation of a text some suggestions regarding the correct rendition of certain words appeared in a message box requiring confirmation on the part of the user. This is another reason of increasing the time the system needed for the translation process which has a negative effect on the computational quality and the overall quality of the system. However, this takes only few seconds and the speed of translation of NA system is considered very good, in general. Such a pause has significance, in particular, for evaluation purposes especially these involving the comparison of systems.

On the basis of the evaluation results of this criterion (See 5.1.1.2.1 above); it has been proved that NA has the lowest performance; while GW has the highest. This is partly due to the reasons mentioned above. Besides, what may slow down the translation speed of NA system is the extra time it requires in searching and looking for the correct and most appropriate equivalent to the SL meaning and structure relying on its four dictionaries, TM in addition to the user dictionary. GW, on the other hand, performs a translation faster than the other Arabic systems as it has only one dictionary involved (i.e., the basic bilingual dictionary).

5.2.3.2 Limitations of User Dictionaries

Dictionaries are the lexical backbone of any MT system. They contain the bulk of the raw information by which a text is translated. The core dictionary used with an MT system offers generic translations which are intended to be accurate in a broad range of texts. For distinguishing between domain-
specific definitions of a word, various user dictionaries with terminology specific to a subject area can be added.

User dictionaries are specifically designed to improve the quality of translated texts by providing the appropriate word and expression for the vocabulary and language of a specific domain. In many cases, where the system is to be used within a narrow specialization, the user himself must supply some of the technical words and expressions.

In the present work, the researcher has designed user dictionaries for the twelve special-domains involved for each of MA and NA systems. Although these dictionaries enhanced the accuracy of terminology rendition of the two systems (as GW lacks such a characteristic), their design and supply of translations to the specific terminologies while the systems were running was not free of errors. The researcher even encountered certain difficulties in this respect especially with NA system. It is worth stressing, here, that even though most of domain-specific terms and certain expressions have been entered into the user dictionaries of these systems, their influence over the total quality of the translated output has not been proved effective (See point c. and Fig. 5.3 in 5.1.1.1.2 and 5.2.2.1 above). Some problems related to this phenomenon concerning the above-mentioned Arabic MT systems are presented below:

(1) MA System
The following points reflected the limitations of the user dictionary and the preference dictionary of MA system:
1.) Even though some terminologies and expressions are entered and coded into the user dictionaries of the system, it selects other equivalents in the TL that do not fit the context for unknown reasons. For instance, the term
‘flagellate’, ‘سُوطيات’ though exists so in the user dictionary is translated into ‘يجد’, as seen below:

Ex. 123
ST: … when the flagellates are experimentally removed.

(From a biological text)

MA: ... عندما يجد يزل بشكل تجريبي.

2.) Another defect of MA’s user dictionary concerns the word forms. It is realized that if the term or expression entered is in the plural form, the system does not translate it when it recurs in its singular form and vice versa. Thus, it is concluded that any change in subsequent occurrences of a word form results in its mistranslation or untranslation. This is illustrated clearly in the following example as the terminology ‘isomers’, ‘إشباه جزيئية’ is left in its SL form when it recurs in the next sentence as a singular noun.

Ex. 124
SL: … some pairs of isomers rotate the plane of plan-polarized light. One member of the pair (… or D isomer) …, while the other (… or L isomer) ….

(From a chemical text)

MA: بعض أزواج إشباه تدير ، استقطب سفح من سفح ضوءاً. واحد من عضو الزوج (… أو دي) … (isomer) … بينما الآخرين … أو إل (isomer)
3.) Terminologies and expressions should be entered and encoded into the user dictionary of MA system according to their sequence in the sentence or text. Any change in this respect leads to translation errors in word positions. The researcher has taken this into account when designing the user dictionaries in the present evaluation. Nevertheless, a typical example not from the data is given below for the purpose of clarification and confirmation:

Ex. 125
ST: All users of Oxford Dictionary ....

MA : كل قاموس أكسفورد مستخدمي ....

Having entered the word ‘dictionary’ before ‘users’ in the user dictionary of MA system, the above translation results instead of the correct rendition of this noun phrase into:

كل مستخدمي قاموس أكسفورد ...

4.) Another dictionary intended to enhance the accuracy of translation is the preference dictionary. It can be designed by the user after the first translation of the text. It is worth stressing that all the words in this dictionary are randomly selected by the system. The user’s role, here, is only to highlight the most appropriate word’s meaning that fits the context. This takes place after the system’s selection of the first meaning in the first attempt which is inappropriate in most cases. In the biological text, for instance, the word ‘culture’ has the meaning ‘زراعة’. In the preference dictionary, the system chooses the first meaning in the list of alternative meanings ‘ثقافة’ where it
does not fit. Thus, the researcher has highlighted the word ‘زراعة’ among
‘ثقافة’، ‘تراث’، ‘مستتبت’، ‘زراعة’ which is the most appropriate to context and
suggested in the HT. Accordingly, this meaning correctly appears in the
second translation of the text. Despite the positive effect of this dictionary on
the translation quality of the texts, there are certain disadvantages of such a
facility:
(i) In many cases, all the suggested vocabularies given in the list of
alternative meanings do not fit the context or the optimal meaning is not
mentioned. Here, the user has no alternative, but to let the system select the
first meaning in the list as it often does. As a result, the SL word will be
mistranslated or deformed in the MT output. In the following extract the
optimal meaning of the word ‘forfeit’, on the basis of the HT, is ‘... 
’ وساؤسق fatt. The other potential meanings given in the preference dictionary are
‘تخسر’، ‘تغزيم’، ‘تتصادر’, where only the latter ‘تخسر’ fits to a certain degree.
Thus, the meaning intended is supplied by the user dictionary thereafter.

Ex. 126
ST: … and shall forfeit my right to any claim or objection.
(From a legal text)

HT :
... وساؤسق حق من أي طلب أو اعتراض.

MA: ... وساؤسق حق إلى أي طلب أو اعتراض.
ii) Some functional words (e.g., both, in, through, etc.) that do not affect much the accuracy of the overall meaning of the sentence or text are randomly selected by the preference dictionary. This is to assign them the exact meaning in the TL. Such an option, in fact, should rather be restricted to the important key words, i.e., content words and terminologies. But, it is found that although such words directly provide the exact meaning and effective conveyance of information, they are mostly not selected.

(2) NA System

As it has previously been mentioned (See 5.2.3.1 above), since the system under evaluation is not the original version of NA, some difficulties during the running of the system and design of its user dictionaries have been encountered. These mainly include the following points:

a. There are few buttons in the user dictionary dialogue box such as: ‘open’, ‘close’, ‘add’, ‘delete’, ‘modify’, ‘ok’ and ‘cancel’ which are essential for its adequate operation and performance. These commands do not appear on the buttons at all and it is really so difficult to do any step without knowing which button is being pressed and for what purpose. With the help of the manual of the system and after many attempts, the researcher could at last realize which button represents what command. The same thing can be applied to the idiom and specific dictionaries (there was no need to design these two dictionaries in this study).

b. In many cases, either when a new word was entered into the user dictionary or when pressing a specific button, a message box appeared indicating that “This programme has performed an illegal operation and will be shut down” and another one pointed out that “If the problem persists,
contact the program vendor.” Consequently, the system closed and required to be restarted again.

c. In general, the design of NA user dictionary is more complex than the same process with MA system. While such an operation requires solely the listing of terms and expressions needed with MA, more information regarding the semantic features (e.g., human, animal, concrete, etc) ; the main grammatical categories of the content words (verb, noun, adjective and adverb) or terms such as: number (singular/plural); gender (Masculine/feminine); type (expression, proper noun, title, etc.); verb (without object, one object, two objects, etc.) are so essential with NA system. Needless to say that such a process is time consuming and requires special efforts without the appearance of commands on the buttons. Above all, the designer of such dictionaries must be highly competent in the linguistic systems of the two languages involved.

d. When typing any word class in the dialogue box, it does not appear; although it is added to the list of words in NA user dictionary. This is confirmed by a second attempt to enter the same word, when a list of items alphabetically arranged displays showing this word highlighted.

e. If the word entered into the user dictionary is capitalized, NA does not translate it when it recurs with a small letter in a particular text.

f. On certain occasions, NA does not translate or mistranslates the words or terms for unknown reasons. It does not take the correct and appropriate translation of these words and expressions from the user dictionary. This is illustrated in the extracts below:

Ex. 127
ST: … the chromosomes begin to coil, **contract** and condense, but the two subunits….  

(From a medical text)

NA:  

... **subunits** ...  

Ex. 128  

ST: … assessment and **restrike** capabilities .  

(From a military text)

... تقييم وريستريكي مؤهلات.

Even though the underlined words have been entered into the user dictionary with their adequate Arabic translations (according to the HT), NA system does not take this into consideration. As a result, it produces deformed and mistranslated words ‘**تنوكوش’ ، ينكمش’ instead of ‘**تنوكوش’ ، ينكمش’، untranslated word ‘**subunits’ which stands for ‘**وحدتين الثانويتين’ and transliterated word ‘**ريستريكي’ instead of ‘**اعادة الضربة’، for unknown reasons.

**g.** Another serious defect in the user dictionary of NA system regards the selection of the optimal and most appropriate meaning to a word or expression. Here, it is necessary to highlight the intended meaning among the others in the long list, as seen below with the alternative meanings the system displays for the word ‘aspect’:

Aspect  

**وجهه، وجهات**
On certain occasions, all the meanings in the list do not fit the context. Thus, the researcher has entered and highlighted a new meaning to the list. As it is mostly the case that the system selects the first meaning because of some functional defects, this requires the deletion of all the alternative meanings. In addition, it is noticeable that when highlighting the intended meaning in the list, it does not appear in the second translation of the text. Again, this
requires the deletion of all the alternative meanings keeping the one desired only; so that the system can select none but this in the second translation.

**h.** In other cases, the alternative meanings of verbs or other lexical items do not appear in the meaning box, but when putting the cursor on different positions in the box only the highlights show. As it is impossible to see and know the meanings hidden, the researcher needs to delete all the highlights and enter the word intended instead.

For all these deficiencies, NA system does not operate adequately and much time and efforts are needed to perform certain functions that are essential to improve the quality of its MT output.

Nevertheless, NA system is found better than MA in the accuracy of the grammatical structure of words and expressions. In other words, the items entered into its user dictionary change according to the grammatical categories and contextual requirements. Further, when a word is added to NA user dictionary, it appears in the same form or in different forms wherever it occurs in the same text.

In general, the effect of the user dictionary on the accuracy of translation and improvement of the MT quality is apparent in the translations of MA and NA systems. A typical example is the following sentence translated by the three participating Arabic systems. Note the poor rendition offered by GW system (where there is no user dictionary) in comparison to those of the other MT systems.

Ex. 129

**ST:** Diagrams showing the refraction of light by an equiconvex lens and by an equiconcave lens are given in Fig.4A.

(From a physical text)
5.3 Summary

In this chapter, results derived from the analysis of texts representing the MT output of the three Arabic systems in addition to some observations regarding the systems’ performance and their computational capabilities have been introduced and discussed. The analysis and discussion rest on three major types of problems. One is related to the understandability and fluency of the MT output, beside its faithfulness to the SL text. The second deals with the most important linguistic criteria determining the wellformedness of the MT output. The third is concerned with some operational problems of systems’ performance during the process of translation.
Results of the first type of problems indicate that the Arabic MT output is, in general, unclear and influent. It also suffers from being unfaithful in the conveyance of SL information. This is due to several factors including grotesque word order and choice, clumsy translation, unknown and mistranslated critical words, partially correct translation, adherence to SL structure and loss and distortion of information.

The major linguistic problems in the data centre around inadequate terminology interpretation, incorrect rendition of SL grammatical structure and erroneously recognized TL morphology.

Further, the operational problems are attributed to certain impediments in measuring the speed of translation and some limitations relevant to the design and performance of the user dictionaries of these MT systems.

A discussion of the main findings of this study and identification of the possible sources of inefficient and ineffective MT quality and performance, are now in order. This will be the task of Chapter Six.

Notes to Chapter Five

(1) For limits of space and for clarity of graph items, reference to criteria types and text types in the figures of the present chapter is not done on the basis of their syllabic division.

(2) Average, here, refers to the overall GSQ of an MT system which is in value between 50-60%, i.e., it indicates ‘fair’ quality of an MT system.
CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

With the application of modern ideas of AI, the field of NLP, in general, and MT, in particular, currently is gaining momentum. MT products are now receiving a considerable amount of interest and many systems are intended to be used by humans to assist them in performing some translation tasks or achieving certain goals. A great deal of activities around MT these days is
evaluation initiative. Their object is an individual or comparative assessment of system performance (in several, not only linguistic respects) and of the quality of linguistic output. This reflects how well the MT system performs the requirements for which it was designed. However, part of the purpose of evaluation is to informatively point out where designs fail to fill requirements.

This study has investigated some functional and non-functional aspects of GSQ to obtain an accurate picture of the capabilities of the currently available Arabic MT systems. Analysis of (268) sentences for each of the selected quality characteristics and MT system representing twelve different text types, has identified and confirmed the existence of various systems’ deficiencies in the translation of texts.

The aim of this chapter is to draw some conclusions on the overall quality assessment of the three Arabic MT systems considered. A summary of the problematic issues and difficulties such systems encounter in the translation process and the possible sources of these problems is also given. On the basis of this, some recommendations toward the development of Arabic MT and systems design will be outlined.

Thus, results obtained in this study support the following conclusions by means of which most of the hypotheses stated in the first chapter have been verified.

6.1 Current State of Arabic MT Systems

6.1.1 General Conclusion

The recent release of a new Arabic MT system for advanced levels, i.e., GW 1.00 (2002) and the new version of MA 3.00 (2002) for professionals, in
addition to the availability of NA 2.00 (2000) in the local market (which has not been subject to any evaluation study before) have encouraged a systematic and standard evaluation of these systems seeking for signs of improvement in Arabic MT. Thus, the present evaluation has been carried out and the results thereof have confirmed that there are still serious drawbacks with these systems.

In general, all the three Arabic MT systems evaluated produce average or below average quality, i.e., with a total quality performance of 58% for MA, 45% for NA and 44% for GW. In fact, their outputs reflect many deficiencies in translating various text types. MA is better than the other systems in performing scientific translations and the three systems can only manage commercial and computational texts with a quite satisfactory translation. The end-user can use them to grasp the general idea of the ST, or translate short, simple texts. Long and complex sentences are especially hard to understand when translated by these systems. One can hardly say that any of them is much better than the others when translating from English into Arabic, as much time and efforts are required for intensive post-editing of their product. In short, they all need serious improvements.

6.1.2 Strengths and Weaknesses of the MT Systems Evaluated

This section provides the findings of the various evaluation types performed to clear up some functional and non-functional abilities of the three Arabic MT systems. Reflecting the merits and demerits of each system in relation to the translation process, such results will be of benefit to the designers and developers of these systems for improvements. Further, they will also be of benefit to the users of these systems to enable them to select the software that can best fulfil their aims and requirements.
A. MA System

MA system has obtained the highest quality as compared to the other systems, though it is still an average quality (i.e., 58%). The TT produced by this system is almost clear, i.e., it is rather difficult to understand the translation of some text types but this is possible after reading the text for two or three times. As a result of making user and preference dictionaries and selecting the theme of translation, the general idea of the text is there and the translation produced by this system is rather informative. It also reflects a fairly faithful translation to the ST. However, due to the large number of grammatical and lexical mistakes, MA gives ill-formed sentences. This has also a bad effect on the adequacy of the domain-specific terminologies included in the user dictionaries of the system. In general, the translation of this system is fairly accurate and can be used as a source of information especially in the fields of computer science, water engineering, petroleum and law. It is the best among the other Arabic MT systems in scientific translation and in its total functional performance, though it should be further improved.

On the other hand, MA system has the best computational performance among the other systems except for the speed of translation as GW is faster, though the difference is not big here. It can translate large texts with typical system storage of 40 Kb in the English and Arabic text boxes. It can also translate large documents of whatever size in the background (e.g., 5.84 Mb/1.165.823 Words in 55 minutes). MA has performed the adaptability and flexibility of 100% as it runs under any version of the operating system and can easily be installed on any PC of whatever specifications. As it has the highest security, it can be registered on one computer only. This system
shows high flexibility with respect to user and preference dictionaries building and updating of its general and special-domain dictionaries with new words and expressions.

**B. NA System**

This Arabic MT system comes second in rank after MA with a total system performance of 45% which is below average (i.e., a less acceptable quality). As a copy and not the original version of the system available, the user suffers from many problems during the run of the system and the translation process. Only with certain text types (e.g., social, legal and commercial), the system can produce average translation where useful information can be extracted. Otherwise, the system performs poorly especially with scientific texts. The texts are hard to read, some meaning of sentences can be gleaned with some effort, reflect poor fidelity to the ST and are ill-formed. The general idea is comprehensible, but it is very hard to read large fragments of the MT product. This is attributed to various kinds of mistakes. In short, the translation of this system can only be used by the end user for grasping the general idea of the ST. as the manufacturers of this system (i.e., CIMOS) have recently released a new version of the system to the global market, it is hoped that some improvements have been made.

It is worth stressing, here, that building the user dictionaries of NA is not an easy task at all. In addition to some operational deficiencies, the process requires high linguistic competence in the two languages involved, i.e., English and Arabic. Thus, the system can better be used by translators or knowledgeable users.

Among the other Arabic MT systems, NA has the lowest non-functional performance. Although it has translated most of the sample texts in this
study in few seconds, its speed is far behind that of GW and MA systems. It suffers from many pauses during the translation process and it cannot translate long and grammatically complex sentences. The largest size of a text it can translate should not exceed 29 Kb, and the system is unable to translate large documents in the background. It is unadaptable to the latest versions of the operating system (i.e., Windows 2000 and XP), but it can easily be installed and run on any PC of whatever specifications. Since it is a copy, it has a zero security and the same copy of the system can be used with all PCs. Further, the flexibility of this MT system is restricted to building user dictionaries.

C. GW System

The evaluation of the first version of GW system which is supposed to be used for advanced level to perform accurate text translation especially in the scientific and engineering domains, has given poor results. Although the quality level of GW is, generally speaking, close to that of NA system, i.e., 44%, it has the worst functional performance which is far behind average for most of the text types except the domain of commerce where it has achieved best (with an average value of 60%). The system can also produce average translation with the social and computer science domains. GW suffers from serious grammatical and lexical problems especially because of the lack of potentiality to design user dictionaries. In general, the ‘suitability’, ‘accuracy’, and ‘wellformedness’ of the texts it produces are poor and far behind average quality. Serious improvements are needed with this system.

With respect to the non-functional quality characteristics, GW is superior to NA system. Like MA system, GW has the highest adaptability and capacity of system storage. It is the fastest Arabic system in performing the
translation, since no extra time is required in search of the domain-specific terminologies or the most appropriate alternative meanings that go with the theme of the text, as it the case with the other MT systems. It has the same average flexibility as NA system, but it is inflexible to dictionaries building.

6.1.3 Other Findings of the Present Evaluation

Other findings of the present evaluation can be summarized by the following points:

1. For the evaluation of ‘readability’, ‘fidelity’ and ‘syntax’, usual methods of rating the quality of MT output have relied on human judges assigning scores on a graded scale (such as 0-3 and 1-5). Although this scoring technique is rather subjective and is itself inconsistent and not very reliable, automatic MTE measures are even less reliable and are still very far from being able to replace human judgment. Nonetheless, the present study has shown that MT can be evaluated using well-known evaluation measures and the metrics and methods adopted have been proved significantly reliable and are the most commonly used in the MTE literature (based on the ISLE’s most recent Evaluation Framework, i.e., EFMTI of 2003 and the researcher’s suggested methods). The approach presented is, indeed, the right way to move towards a trustworthy systematic evaluation of MT quality. In general, human judgment still seems to be the most popular way of assessing quality characteristics.

2. To obtain reliable results of evaluation, the correlation coefficient values between the scorings of the two evaluators for the criteria of ‘readability’, ‘fidelity’ and ‘syntax’ are above 0.9 which indicate a statistically significant correlation.
3. The GSQ for the three Arabic MT systems is found. This is based on the results of the functional and non-functional criteria. The highest system’s total performance is that of MA with a value of 58% for GSQ. The values for the other systems are 45% for NA and 44% for GW.

4. The highest functional performance is obtained by MA with a value of 40.15 out of 80 and the lowest value of such a performance is scored by GW system with a value of 28.5 out of 80.

5. The highest non-functional performance is achieved by MA system with a value of 17.85 out of 20; while the lowest value of 10.65 out of 20 is given to NA in this respect.

6. Each of the functional criteria in the present study has been calculated in terms of text types. The GSQ of each of the systems evaluated is also determined in this respect. The results of such tests are graphically represented. It has been found that the best system’s value for MA is in the field of computer science (69%) while the lowest value is in the domains of ‘biology’ and ‘chemistry’ (52%). As for NA system, the highest value is in the field of ‘social science’ (53%); however, the lowest performance is with the military text (34%). GW system performs best with the commercial / business domain (60%) whereas it has the lowest performance with the chemical text (30%).

7. The correlation coefficients between the five functional criteria for each Arabic MT system are tested using cross correlation matrix. This is to statistically determine the degree of interrelation among these criteria since they are linguistically highly interrelated. The results are tabulated showing that for GW, 88% of the r values are significant indicating that the data of the interrelation among the functional
criteria fit with Person’s model and are of linear relation. However, for both MA and NA systems, this test points out to a relationship between two sets of variables, but it is not linear. The aim of these tests is only to show the possibility of carrying out statistical studies on the different evaluation results.

8. Three major types of problems have been identified in the data of the Arabic MT systems, which are the main causes of poor quality. These are cognitive, linguistic and operational problems. Indeed, linguistic phenomena are particularly troublesome and are the direct factors of cognitive problems.

9. Cognitive problems reflecting unclear and inaccurate MT output and unfaithfulness to the ST result from a variety of phenomena most of which are linguistic ones. These include:
   a.) Grotesque word order and choice.
   b.) Understandable clumsy translation.
   c.) Unknown and mistranslated critical words.
   d.) Partially correct translation.
   e.) Adherence to SL structures.
   f.) Loss and distortion of information.

10. The texts of the data have revealed a number of linguistic problems peculiar to the grammatical structure of words and sentences and domain-specific terminology. Such deficiencies in the Arabic MT output are due to:
   1. Inadequate terminology interpretation.
   2. Incorrect rendition of SL grammatical structures.
   3. Erroneously recognized TL morphology.
11. Various types of lexical and semantic problems have been identified in the data of the Arabic MT systems. These are direct causes of inadequate terminology rendition, involving:
   a.) Untranslated terminology.
   b.) Forms irrelevant to the TL.
   c.) Synonym usage.
   d.) Inappropriate word choice.
   e.) Mistranslations.
   f.) Wrong ordering of phrasal constituents of terms.
   g.) Errors of Category and Word Class.

12. There are major syntactic deficiencies in the data of the three Arabic MT systems that significantly alter structure and meaning and badly affect the fidelity of the MT output to the original ST and the understandability of the TT. The results of the present evaluation confirm that none of the Arabic MT systems is good at rendering English syntax correctly. It seems that more complex restructuring may well be beyond the capacity of the Arabic systems. The major system’s handicap in Arabic MT is that of word order. In addition, some other syntactic problems that deteriorate the quality of the Arabic MT output are:
   a.) Bizarre Syntactic arrangement.
   b.) Lack of grammatical arrangement.
   c.) Differences in case.
   d.) Incorrect pronoun translation.
   e.) Wrong part-of-speech categorization.
   f.) Inconsistent preposition handling.
   g.) Failure to treat the definite article.
h.) Problems of tense, aspect and voice.

13. Another serious and very complex linguistic problem that results in generating ill-formed Arabic MT output, is that of incorrect Arabic morphological structure. In fact, an important consideration in the development of Arabic MT systems is the matter of Arabic morphology. It is found that all the evaluated Arabic MT systems exhibit uniformly low morphological representations. They are incapable of recognizing the morphological differences and their equivalent structures between English and Arabic. For these and other sources of problems, some morphological errors have been identified in the data of the present study. These directly result from certain lexical and syntactic phenomenon and concern mainly the deficiencies of affixes. The morphological errors found in the data include:
   a.) Inappropriate choice of prefixes.
   b.) Errors of word infixes.
   c.) Inability to maintain correct suffixes.

14. The present evaluation indicates that both GW and MA systems, produced by the same company (ATA), use the same parsing engine as they perform almost the same for all the grammatical categories. The superiority of MA quality over that of GW with respect to the other functional criteria is, in fact, due to the positive effect of the user dictionaries and the potentiality of selecting the theme of translation that facilitates the choice of the appropriate lexical item that fits the context during the translation process. This helps to improve the grammatical, lexical and semantic aspects of the TT.

15. Two types of operational problems have been recognized during the run of the Arabic MT systems and the translation process. These concern the
measurement of the translation speed and the design of the user
dictionaries:

I.) Some errors hinder the measurement of translation speed with NA
system. Accordingly, the system does not work adequately and
suffers from constant pauses that impede the translation of texts and
increase the time required for this process. These defects are the
major cause of obtaining the lowest performance of speed among the
other Arabic MT systems.

II.) Regarding MA and NA systems, special-
domain user dictionaries
are made by the researcher to enhance the accuracy of terminology
rendition. Nevertheless, the design and supply of translations and the
effectiveness of the terms and expressions in the TT have not been
free of errors. This is attributed to certain problems with each of
these systems:

III.) Even though some terminologies and expressions are entered and
encoded into the user dictionaries of MA, they are translated
inadequately because:

a. the system selects other equivalents in the TL that do not fit
the context,

b. there is a change in the word form for grammatical purposes
which differs from the one entered and encoded into the
dictionary and,

c. there is a change in the sequence of the terminology or
expression entered and its position in the sentence or text.

Other problems with this system concern the preference
dictionary. These include:
a. The suggested vocabularies in the list of alternative meanings either do not fit the context or the optimal meaning is not found.

b. Most of the key words, useful expressions and terminologies are not selected by this dictionary where other functional words (e.g., articles or prepositions) that do not have much significance in the text are included.

Some problems encounter the user of NA system with respect to the design and running of the domain-specific user dictionaries. These are mainly due to the fact that the system currently available in the local market is a copy and not the original version. The major difficulties with this Arabic MT system concern:

a. The disappearance of the commands on the buttons essential for the adequate running and performance of these dictionaries.

b. Constant shut downs of the system for the performance of illegal operations or certain errors that appear in a message box.

c. The complexity of designing user dictionaries of this system requiring, in addition to high competence in the linguistic systems of the languages involved, much time and efforts using buttons without commands.

d. Entering words, expressions, idioms and terminologies without being able to see them, only after a second attempt and following special steps to be discovered by the user.

e. Untranslation of words and expressions if they are entered with capital letters and vice versa.
f. For unknown reasons, NA system, sometimes, mistranslates words or expressions that occur correctly in its user dictionary.

g. The complexities accompanying the assignment of the optimal meaning in the list of alternative meanings.

Despite all these inadequacies, the user dictionaries of MA and NA have been proved effective in the present evaluation. This is apparent in the results of the GSQ assessment and the total performance of the functional criteria where both MA and NA are superior to GW system. Besides, the accuracy of the grammatical structure of words and expressions of NA user dictionaries have been found better than these of MA on the basis of the analysis of Arabic MT output in the data.

16. The current Arabic MT systems can best be used as a ‘gisting’ application, although this does not deliver the quality that human translators are capable of producing. As they are very fast in performing the translation (i.e., more than 10,000 w/min), they can rapidly provide the gist of large texts with low-cost translation. This, however, works only with certain text types.

17. As the Arabic MT output reflects disappointing quality in most cases, much time and considerable efforts are required to perform intensive post-editing to the various types of errors it contains.

6.2 Some Sources of Arabic MT and MTE Problems

The discussion in Chapter Five and the conclusions in section 6.1, point out to serious shortcomings in the quality of Arabic MT output and systems’ performance. This section attempts to explore some possible sources of
problem regarding the poor quality performance of Arabic MT systems. More extended investigation is required to pin-point other sources of problem.

6.2.1 Variations between SL and TL Linguistic Structures
Due to the large differences between the structures of languages from completely different linguistic families such as Arabic and English, and to related cultural differences, translation requires the use of entirely different linguistic expressions in the two languages. The differences in syntax and semantics between Arabic and English are simply too great.

This, in fact, is one of the major problems MT into Arabic has encountered. Moreover, certain aspects like the rich Arabic vocabularies, alternative meanings and equivalents and the many English words that have more than one meaning in Arabic constitute a huge number of obstacles for Arabic MT. Further, the sheer number and complexity of the Arabic linguistic rules makes programming of MT systems nightmarish. Thus, the elucidation of translation techniques and requirements which take such factors into account is a major undertaking.

6.2.2 Deficiencies of Systems’ Dictionaries and Components
Arabic MT is a software rather than a hardware problem. The various Arabic MT problems in the data are attributed to certain deficiencies in the systems’ linguistic components and transfer rules. Therefore, further developments and improvements are required. In addition, the special-domain dictionaries which include many terms and expressions of specialized domains can enhance the quality of translation of these subjects. Unfortunately, they do not run during the translation process and are there only for further
information or post-editing process. The lack of the potentiality of designing user dictionaries, as it is the case with GW system, also degrades the translation quality of MT systems. This problem results in the existence of untranslated and mistranslated terminologies and key words, which is generally a great source of semantic errors and wrong syntactic categorization.

MT is determined above all by the quality and range of its dictionary information. The present evaluation indicates that the dictionaries of the Arabic MT systems still need a wealth of information in their entries including part of speech, grammatical categories, morphological information, semantic features, etc. Moreover, in these dictionaries there are missing information regarding coverage of word sense and knowledge of the world. They provide hardly enough contextual information too. This eventually leads to low system’s performance and decreases translation accuracy.

**6.2.3 The Frail Nature of MT**

The principal impediment to MT, including Arabic MT of course, is that the computer or the computer programme does not understand an ST at least not in the human sense of the word. There is a limited capacity and intelligence of computers and MT systems. In understanding a linguistic utterance in context, three interactive levels of comprehension must be distinguished: 1.) syntactic, 2.) lexical and 3.) pragmatic understanding. Of these three levels of understanding, computer programmes can handle syntactic properties to a considerable degree, semantic properties to some degree, and pragmatic issues which are related to knowledge of the world to zero degree. Many AI and MT researches have been carried out in order to change this situation.
The existing Arabic MT systems are meant for a variety of scientific, legal, commercial, engineering and social sciences. Early researches indicate that under controlled conditions and by using pre-tested texts, MT output is only 80% accurate with abysmal syntax and style. Thus, the systems evaluated in this study should be limited to very restricted domains, i.e., not to be too general, to get good-quality translation. This will limit the vocabulary and somewhat the grammar making the problem of translation acceptably feasible for computer programming.

6.2.4 Lacking Awareness of the Strategic Nature of Arabic MT
Arabs should realize the importance of MT in relation to their language and their culture. This is in order to enable Arabs to keep up with the advancements in the world and in catching up with the rest of the world at all levels. Unfortunately, efforts in Arabic MT within or outside the Arab world remain minimal and Arabic MT products remain in their infancy. This is because human translation from and to Arabic is by far less institutionalized than its counterparts in the West, and such an awareness of the crucial and strategic nature of MT is still lacking in the Arab world. Given that non-Arabic MT has been recognized very early as a strategic area, and given that funding has poured in to help its development, the discipline has been active for half a century already.

6.2.5 Lacking Awareness of the Significance of MTE
For the disappointing results obtained from studies such as the present one, reflecting the poor quality of output frequently produced by Arabic MT systems; the manufacturers should realize the importance of evaluation research before they release such products to the market. In fact, evaluation
is one of the most important stages in the life of an MT system. If ignored, and not carried out scientifically and objectively, the quality of the MT software will not improve. Like the case with the Arabic MT systems, Arabic MTE is far behind the recent advancements in the world with respect to the various researches, conferences and workshops on MT evaluation. There is a big gap represented by the lack of awareness of the great world contributions and significance of this area of research and development. As a result of rare Arabic MTE studies and lack of experts and researchers in this field, no significant improvement is noticeable in the quality of the recently released new Arabic MT systems to the local and global market.

In addition to the above-mentioned sources of problems, it is noteworthy to indicate that all efforts in the field of Arabic CL and MT imitate the approaches and technology utilized in the processing of English. This leads to ignore the essential requirements of the Arabic end-user and the characteristic features of Arabic.

All in all, the researcher, in this section, has outlined the possible sources of problems faced by designers and users of Arabic MT systems, and why these problems are relatively easy or difficult to tackle. Some difficult problems may prove to be inherently unsolvable. Some are certainly intractable with present methods and at the present stage of knowledge. For others, there are good prospects of viable approaches; research continues and it is hoped that gradual if not dramatic improvements will take place in the future.

6.3 Recommendations
For the purpose of getting Arabic MT and MTE to the stage which MT for other languages has reached, the following recommendations are suggested:

1. **Awareness of the Strategic Nature of Arabic MT**

   The emergence, in recent years, of international networks of translation companies and of translation companies specialized in one or more domains is a strategic response to changes in the pattern and nature of translation demand. These developments should encourage Arab translators to make greater use of translation tools and translator workstations. It is essential that the Arab public, governmental institutions, private companies, universities and research centers get involved in supporting basic research and development of Arabic MT tools and MTE studies. Increasing awareness of the strategic nature of Arabic MT and MTE will be the driving force that will encourage and fund research groups and projects for improving the currently existing MT systems and other translation tools, and for developing this area, in general.

2. **Staff Development Schemes**

   What is also required is a staff development scheme whereby a group of programmers, linguists, translators and computer specialists who are interested in MT should be motivated and involved in intensive training programmes to acquire the advanced skills and knowledge in the field of MTE. They should take part in local and international workshops and seminars, participate in special training courses on modern and effective techniques of using evaluation frameworks, metrics, methods and other skills, and attend symposiums and conferences, exhibitions and trade shows on topics of a similar nature. Such schemes, if well organized, will
immediately give returns in terms of improvement and development of Arabic MT and systems’ design and quality.

3. Teaching MTE
In the Departments of Computer Science, Linguistics and Translation at Arab Universities, more modern methods and classroom techniques which emphasize the advantages of MT and the usefulness of MTE should be used. The development of learner’s skills and the knowledge they gain regarding the principles of MTE, types of evaluation, the use of metrics and statistical methods and why, by whom and for whom evaluation may be carried out, is so essential to prepare experts and advanced cadre in this respect. In addition, a variety of recent publications and other custom made material should be made available to teachers and learners as well, to serve their needs and promote their motivation.

4. Cooperation among Arabic MT Companies
A storage relationship and cooperation must be built among the Arabic MT companies to work on problems of common interest and direct their efforts and expertise towards one goal, i.e., the improvement and development of Arabic MT. Team work of designers, developers and researchers representing these companies is badly needed where they can discuss their findings, their products and problems they face seeking for a better Arabic MT. Concentration should be on glass-box evaluation approach to overcome shortcomings and problems of Arabic MT systems’ components and transfer tools. Monitoring and evaluation of existing Arabic MT systems and identification of their shortcomings is necessary and should be done on a
continuous basis. Experts from non-Arab countries, who have a proven record in the field of MT and MTE can help in this respect too.

5. Other Recommendations

To achieve a better quality of Arabic MT systems and secure advancement of Arabic MT, the following requirements have to be met:

1.) Governments and Pan Arab organizations should give Arabic MT a boost establishing CL centers to enrich Arabic research in NLP and Arabic MT theory and practice.

2.) As a research area in itself with its own conferences and literature, evaluation of MT should concentrate on the evaluation and improvement of the lexicons of such systems. Domain-specific dictionaries should run during the translation process by making a special code for each domain and text type to precisely select the meaning of terms.

3.) Contextual information, knowledge of the world, a broad coverage of word senses are essentially to be encoded in a system’s dictionary. Other grammatical, pragmatic and semantic aspects should also be taken into account.

4.) There is a need for continuing efforts to improve quality where ‘close’ translation is desired (e.g., the development of subject-specific systems). Systems that are restricted to special text types and domains and are limited to specific purposes can produce high quality MT output with a minimum of error and a maximum of precision and speed. This can be said about using texts with a controlled language, as input data, to guarantee a smooth and unproblematic translation.
5.) Concentrating on the production of interactive MT systems where a human translator can interact with the computer and make the necessary changes and corrections during the translation process for a better quality of output. Further, the compilation of extensive knowledge bases such as dictionaries, glossaries and TMs will help improve MT.

6.) Emphasizing the advantages of pre-editing and post-editing processes on the quality of an MT output is a necessity. In fact, MT and post-editing have recently become viable solutions to meet the growing translation demand. Thus, formal training in post-editing techniques and skills is so essential to many translation professionals.

7.) For ATA company that has produced GW and MA systems, the results of the present evaluation suggest that instead of spending much time and efforts on the production of MT systems with poor quality (e.g., GW) just for purely commercial and financial benefits; this company can cancel the ineffective MT system (which badly affects the confidence of its customers) and concentrate on the improvement and development of more promising Arabic MT systems like MA.

6.4 Suggested for Further Research

As to the implications for research, this study seems to indicate various directions for further research. Other types of evaluation studies, addressing different criteria, and adopting different methods, metrics and other text types need to be carried out.

In short, a great deal of research in the field of MTE is essential to discover all the capabilities of the currently existing Arabic MT systems, their shortcomings and potentiality of improvements. What the findings of the present work suggest is that a further study is necessary.
REFERENCES


___________.(2002b). “Types of MTE.” E-mail to the researcher. 13 July 2002 b.
___________.(2002d). “Sample Text Types and Evaluation Methodology.” E-mail to the researcher 25 Aug. 2002 d.
___________.(2002e). “Metrics and scoring schemes.” E-mail to the researcher. 31 Aug. 2002 e.
___________.(2002f). “GSQ Weighting Scheme.” E-mail to the researcher. 30 Sept. 2002.


__________. (2002 a.) “ISLE Taxonomy.” E-mail to the researcher. 25 June 2002a.

__________. (2002 b). “Sample Texts for MTE.” E-mail to the researcher. 4 Sept. 2002 b.

__________. (2002c). “GSQ Weighting Scheme.” E-mail to the researcher. 23 Sept. 2002 c.


__________. (2002e). “Scientifically Justifiable Weighting Scheme.” E-mail to the researcher. 2 Oct. 2002 e.

__________. (2002f). “Overall Summed Score.” E-mail to the researcher. 3 Oct. 2002 f.


**English Data**


Arabic Data


عبد الرحمن علوش .(سنة النشر غير معروفة).الطريقة الميسرة للمراسلات التجارية باللغة الانجليزية.( مكان النشر غير معروف) دار الشمال للطباعة و النشر و التوزيع.

المركز العربي للتعليم والترجمة والتآليف والنشر.
English – into – Arabic MT Systems


APPENDIX A

FEMTI - a Framework for the Evaluation of Machine Translation in ISLE

EXPANDED CLASSIFICATION

1 Evaluation requirements

- 1.1 Purpose of evaluation
  - 1.1.1 Feasibility evaluation
  - 1.1.2 Requirements elicitation
  - 1.1.3 Internal evaluation
  - 1.1.4 Diagnostic evaluation
  - 1.1.5 Declarative evaluation
  - 1.1.6 Operational evaluation
  - 1.1.7 Usability evaluation

- 1.2 The object of evaluation
  - 1.2.1 A component of an MT system
  - 1.2.2 An MT system considered as a whole
  - 1.2.3 An MT system considered as a component of a larger system

- 1.3 Characteristics of the translation task
  - 1.3.1 Assimilation
    - 1.3.1.1 Document routing / sorting
    - 1.3.1.2 Information extraction / summarization
    - 1.3.1.3 Search
  - 1.3.2 Dissemination
1.3.2.1 Internal / in-house publication
   - 1.3.2.1.1 Routine
   - 1.3.2.1.2 Experimental / research

1.3.2.2 External publication
   - 1.3.2.2.1 Single-client
   - 1.3.2.2.2 Multi-client

   1.3.3 Communication
     - 1.3.3.1 Synchronous
     - 1.3.3.2 Asynchronous

1.4 User characteristics
   - 1.4.1 Machine translation user
     - 1.4.1.1 Education
     - 1.4.1.2 Proficiency in source language
     - 1.4.1.3 Proficiency in target language
     - 1.4.1.4 Computer literacy
   - 1.4.2 Translation consumer
     - 1.4.2.1 Proficiency in source language
     - 1.4.2.2 Proficiency in target language
   - 1.4.3 Organisational user
     - 1.4.3.1 Quantity of translation
     - 1.4.3.2 Number of personnel
     - 1.4.3.3 Time allowed for translation

1.5 Input characteristics (author and text)
   - 1.5.1 Document type
     - 1.5.1.1 Genre
     - 1.5.1.2 Domain/field of application
   - 1.5.2 Author characteristics
1.5.2.1 Proficiency in source language
1.5.2.2 Professional training

1.5.3 Characteristics related to sources of error
1.5.3.1 Intentional error sources
1.5.3.2 Medium related error sources
1.5.3.3 Performance related errors

2 System characteristics to be evaluated

2.1 MT system-specific characteristics

2.1.1 Translation process models
2.1.1.1 Methodology
2.1.1.1.1 Rule-based models
2.1.1.1.2 Statistically-based models
2.1.1.1.3 Example-based models
2.1.1.1.4 Translation memory models
2.1.1.2 Models
2.1.1.2.1 Direct
2.1.1.2.2 Transfer
2.1.1.2.3 Interlingua

2.1.2 Linguistic resources and utilities
2.1.2.1 Languages
2.1.2.2 Dictionaries
2.1.2.3 Wordlists, glossaries
2.1.2.4 Comparable and parallel corpora
2.1.2.5 Grammars

2.1.3 Characteristics of process flow
• 2.1.3.1 Translation preparation activities
• 2.1.3.2 Post-translation activities
• 2.1.3.3 Interactive translation activities
• 2.1.3.4 Dictionary updating
• 2.1.3.5 Process management

• 2.2 System external characteristics
  o 2.2.1 Functionality
    • 2.2.1.1 Suitability
      • 2.2.1.1.1 Target-language only
        • 2.2.1.1.1.1 Readability (or: fluency, intelligibility, clarity)
        • 2.2.1.1.2 Comprehensibility
        • 2.2.1.1.3 Coherence
        • 2.2.1.1.4 Cohesion
      • 2.2.1.1.2 Cross-language / contrastive
        • 2.2.1.1.2.1 Coverage
          • 2.2.1.1.2.1.1 Cross-language phenomena
          • 2.2.1.1.2.1.2 Corpus-based phenomena
        • 2.2.1.1.2.2 Style
    • 2.2.1.2 Accuracy
      • 2.2.1.2.1 Fidelity
      • 2.2.1.2.2 Consistency
      • 2.2.1.2.3 Terminology
    • 2.2.1.3 Wellformedness
      • 2.2.1.3.1 Punctuation
      • 2.2.1.3.2 Lexis / lexical choice
      • 2.2.1.3.3 Grammar / syntax
      • 2.2.1.3.4 Morphology
- 2.2.1.4 Interoperability
- 2.2.1.5 Compliance
- 2.2.1.6 Security

○ 2.2.2 Reliability
- 2.2.2.1 Maturity
- 2.2.2.2 Fault tolerance
- 2.2.2.3 Crashing frequency
- 2.2.2.4 Recoverability
- 2.2.2.5 Reliability compliance

○ 2.2.3 Usability
- 2.2.3.1 Understandability
- 2.2.3.2 Learnability
- 2.2.3.3 Operability
- 2.2.3.4 Documentation
- 2.2.3.5 Attractiveness
- 2.2.3.6 Usability compliance

○ 2.2.4 Efficiency
- 2.2.4.1 Time behaviour
  - 2.2.4.1.1 Pre-processing time
    - 2.2.4.1.1.1 Pre-editing time
    - 2.2.4.1.1.2 Code-set conversion
    - 2.2.4.1.1.3 Preparation time
  - 2.2.4.1.2 Input-to-output translation speed
  - 2.2.4.1.3 Post-processing time
    - 2.2.4.1.3.1 Post-editing time
    - 2.2.4.1.3.2 Code-set conversion
    - 2.2.4.1.3.3 Update time

- 2.2.4.2 Resource utilisation
- 2.2.4.2.1 Memory
- 2.2.4.2.2 Lexicon
- 2.2.4.2.3 Clean-up
- 2.2.4.2.4 Program size

- 2.2.5 Maintainability
  - 2.2.5.1 Analyzability
  - 2.2.5.2 Changeability
    - 2.2.5.2.1 Ease of upgrading multilingual aspects of system
  - 2.2.5.2 Improveability
  - 2.2.5.2.2 Ease of dictionary updating
  - 2.2.5.2.3 Ease of modifying grammar rules
  - 2.2.5.3 Stability
  - 2.2.5.4 Testability
  - 2.2.5.5 Maintainability compliance

- 2.2.6 Portability
  - 2.2.6.1 Adaptability
  - 2.2.6.2 Installability
  - 2.2.6.3 Portability compliance
  - 2.2.6.4 Replaceability
  - 2.2.6.5 Co-existence

- 2.2.7 Cost
  - 2.2.7.1 Introduction cost
  - 2.2.7.2 Maintenance cost
  - 2.2.7.3 Other costs

---

General Software Quality

System External

---

Functional Criteria

Non-Functional Criteria
Figure (3.11). Model of Evaluating the Total Performance (the External Quality of Translation) of Arabic MT Systems.

APPENDIX B

The researcher would like to thank the members of the jury who have so carefully considered the lists of criteria with the appropriate metrics
and methods of evaluating each, in addition to the scoring scheme of finding out the overall value of General Software Quality. The members of the jury are:

1.) Prof. Dr. I’naad Ghazwan / Dept. of Arabic / College of Languages / Univ. of Baghdad.

2.) Prof. Majeed Al-Mashtaa / Dept. of Translation / College of Arts / AL-Mustansiriyya Univ.

3.) Prof. Dr. Sabaah Al-Rawi / Dept. of English/College of Languages/ Univ. of Baghdad.

4.) Prof. Dr. Eduard Hovy / Information Science Institute / Univ. of Southern California.

5.) Prof. Dr. Andrei Popescu-Belis / ISSCO / TIM / ETI / Univ. of Geneva.

6.) Asst. Prof. Dr. Abaas Fadhil Al-Hashimi / Dept. of Post-Graduate Studies / National Center for Computer Science.

7.) Asst. Prof. Dr. Abdul-Latif Al-Jumayli / Dept. of English/ College of Arts/ Univ. of Baghdad.

8.) Asst. Prof. Dr. Sa’ad Abdul-Sataar / Dept. of Post-Graduate Studies / National Center for Computer Science.

9.) Asst. Prof. Dr. Ahmed Guessaum / Dept. of Computer Science / Univ. of Sharjah.

10.) Asst. Prof. Dr. Sawsen Al-Samir / Dept. of English / College of Arts / AL-Mustansiriyya Univ.
To the Jury Members

Dear Mr... /Ms.

The researcher is conducting a Ph.D. project aiming at evaluating English -into- Arabic MT systems with respect to certain criteria of General Software Quality.

To fulfil this aim, the appropriate metrics and methods of evaluating each of the selected quality characteristics and sub-characteristics have been chosen. Moreover, the scoring scheme of finding out the overall value of the system’s total performance has also been determined.

The researcher has the honour of nominating you as one of the jury members to decide on the face and content validity of the above-mentioned aspects of evaluation. Your invaluable suggestions and comments would be highly appreciated.

Thank you very much in advance for your cooperation.

Yours sincerely,

Ph.D. Candidate: Yasmin Hikmet.
The following are definitions of General Software Characteristics and sub-characteristics under evaluation in the present research work, with the suggested methods of evaluating each:

**Functional Attributes:**
1.) **Functionality:** a set of attributes that bear on the existence of a set of functions and their specified properties. The functions are these that satisfy stated or implied needs.

**How to Measure:**
Combination of the scores of the sub-attributes it contains.

1. **Suitability:** attribute of software that bears on the presence and appropriateness of a set of functions for specified tasks.

**How to Measure:**
Combination of the scores of the sub-attributes it contains.

a.) **Readability:** the degree to which some discernible meaning is conveyed by each sentence in a text. For this measure, the sentence neither needs to make sense in the context of the rest of the text, nor be grammatically well-formed since these features would be measured by coherence and syntax.

**How to Measure:**
A 4-point scale.
2. **Accuracy**: attributes of software that bear on the provision of correct agreed upon results or effects.

**How to Measure:**
Combination of scores of the sub-attributes it contains.

a.) **Fidelity**: subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation.

**How to Measure:**
Rating of sentences read in their context on a 4-point scale.

b.) **Terminology**: subjective evaluation of the degree to how correctly the most important terms are translated.

**How to Measure:**
Percentage of input /output domain terms mistranslated.

3. **Wellformedness**: This involves the following:

a.) **Morphology**: Degree to which words are correctly inflected (e.g., tense, number, gender, case, etc.). This is especially important for highly inflected languages like Arabic.
**How to measure:**

Finding out the average value by dividing the number of corrections by the total number of inflectable words per text. The more corrections made, the lower the quality of the text will be.

**b.) Syntax:** Grammatical corrections of the sentence.

**How to Measure:**
Rating of sentences on a 5-point scale.

**Non-Functional Attributes:**

1. **Efficiency:** a set of attributes that bear on the relationship between the level of performance of the software and the amount of resources used under stated conditions.

**How to Measure:**
Combination of scores of sub-attributes it contains.

**a.) Production Time/Speed of Translation:** the time required for translating a particular test text. The translation production time refers to the time between a request for a translation and reception.

**How to Measure:**
Using clock time for measuring the number of words translated per minute and finding the percentage value for every 100 words/min.
2. **Storage**: refers to the capacity of the system with respect to the size it takes of the hard disk (in Kb), i.e., to run the system adequately and the minimum hardware requirements of PC it needs.

**How to measure:**

Dividing the highest storage capacity of an MT system by the typical storage obtained in this study.

3. **Portability**: a set of attributes that bear on the ability of software to be transferred from one environment to another.

**How to Measure:**

Combination of scores of sub-attributes it contains

a.) **Adaptability**: attributes of software that bear on the opportunity for its adaptation to different specified environments without applying actions or means than those provided for this purpose for the software considered.

**How to Measure:**

Using a special checklist of certain values. Then, calculating the percentage value using 5–point scale for this purpose.

1.) **Flexibility/Extensibility**: the modification capacity of the system with respect to the quantities and qualities of additions (e.g., new lexical items to the dictionary of the system) or other similar processes or changes.
**How to Measure:**

Using a special checklist of binary values. Then calculating the percentage values using a 3-point scale for this purpose.

Now, to give values to the mother node, different scores of the sub-attributes are to be combined to give the general score. This relies heavily on the relative importance of the attribute based on the aims and requirements of any particular study or research (EAGLES, 1996).

In the present study, much attention is given to the functional characteristics which are the most critical for any MT evaluation. To achieve objectivity and avoid counter arguments and as the sub-attributes under ‘Suitability’, ‘Accuracy’ and ‘Wellformedness’ are closely related from the translation viewpoint, they have been given equal values with respect to the General Software Quality (GSQ).

The non-functional attributes under evaluation, i.e., ‘Speed of Translation’; ‘Adaptability’ and ‘Flexibility’ are closely related to the MT process and are considered of equal importance.

Accordingly, the distribution of percentage values (average scores) of the ‘Functional’ and ‘Non-functional’ characteristics to the GSQ is shown below:

1- Functional attributes=80% where ‘Accuracy’=32% and ‘Wellformedness’ = 32%, while ‘Suitability’ = 16%. Each of the sub-attributes under ‘Accuracy’ and ‘Wellformedness’ = 16% for the reasons mentioned above.
2- Non-functional attributes = 20% where each of the sub-attributes it contains, i.e., “Efficiency”; ‘Portability’; ‘Storage’ and ‘Flexibility’ = 5%.

### APPENDIX (C)

**Sample Checklist for Adaptability and Flexibility**

**Inspection in the Three MT Systems under Evaluation**

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>CHARACTERISTIC</th>
<th>FEATURE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL-Mutarjim AL-Arabey</td>
<td>Adaptability</td>
<td>Not adaptable, Partially adaptable, Fairly adaptable, Good adaptability, Very good/superior adaptability</td>
<td>0-1-2-3-4</td>
</tr>
<tr>
<td></td>
<td>Flexibility</td>
<td>Very good/Superior flexibility, Limited/Fair flexibility, No flexibility</td>
<td>0-1-2</td>
</tr>
<tr>
<td>Golden AL-Wafi</td>
<td>Adaptability</td>
<td>Not adaptable, Partially adaptable, Fairly adaptable, Good adaptability, Very good/superior adaptability</td>
<td>0-1-2-3-4</td>
</tr>
<tr>
<td></td>
<td>Flexibility</td>
<td>Very good/Superior flexibility, Limited/Fair flexibility, No flexibility</td>
<td>0-1-2</td>
</tr>
<tr>
<td>An-Nakel AL-Arabi</td>
<td>Adaptability</td>
<td>Not adaptable, Partially adaptable, Fairly adaptable, Good adaptability, Very good/superior adaptability</td>
<td>0-1-2-3-4</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Very good/ Superior flexibility, Limited/ Fair flexibility, No flexibility</td>
<td>0-1-2</td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX (D)

The Experimental Values of Texts % Readability

<table>
<thead>
<tr>
<th>Text type</th>
<th>Evaluator</th>
<th>System Type</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Golden Al-Wafi</td>
<td>Al-Muterjim Al-Arabey</td>
<td>An-Nakel Al-Arabi</td>
<td></td>
</tr>
<tr>
<td>Biological</td>
<td>1</td>
<td>34.36</td>
<td>56.06</td>
<td>30.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>30.33</td>
<td>51.51</td>
<td>24.24</td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>1</td>
<td>30.55</td>
<td>48.44</td>
<td>46.61</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>32.66</td>
<td>41.77</td>
<td>50.11</td>
<td></td>
</tr>
<tr>
<td>Computational</td>
<td>1</td>
<td>68.66</td>
<td>81.66</td>
<td>76.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>64.00</td>
<td>81.66</td>
<td>70.00</td>
<td></td>
</tr>
<tr>
<td>Financial</td>
<td>1</td>
<td>36.46</td>
<td>62.82</td>
<td>50.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>40.04</td>
<td>64.10</td>
<td>51.43</td>
<td></td>
</tr>
<tr>
<td>Legal</td>
<td>1</td>
<td>50.00</td>
<td>80.36</td>
<td>59.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>50.21</td>
<td>79.33</td>
<td>53.33</td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>1</td>
<td>40.96</td>
<td>72.27</td>
<td>43.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>42.12</td>
<td>69.57</td>
<td>45.45</td>
<td></td>
</tr>
<tr>
<td>Petroleum</td>
<td>1</td>
<td>32.17</td>
<td>76.16</td>
<td>66.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29.35</td>
<td>80.75</td>
<td>71.35</td>
<td></td>
</tr>
<tr>
<td>Military</td>
<td>1</td>
<td>44.51</td>
<td>62.12</td>
<td>27.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>43.12</td>
<td>64.18</td>
<td>22.18</td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>1</td>
<td>46.06</td>
<td>57.57</td>
<td>39.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>42.51</td>
<td>60.60</td>
<td>41.25</td>
<td></td>
</tr>
</tbody>
</table>
# APPENDIX (E)

## The Experimental Values of Texts % Fidelity

<table>
<thead>
<tr>
<th>Text type</th>
<th>Evaluator</th>
<th>System Type</th>
<th>Golden</th>
<th>Al-Muterjim</th>
<th>An-Nakel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Al-Wafi</td>
<td>Al-Arabey</td>
<td>Al-Arabi</td>
</tr>
<tr>
<td>Biological</td>
<td>1</td>
<td>25.75</td>
<td>36.36</td>
<td>21.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>25.30</td>
<td>35.34</td>
<td>18.24</td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>1</td>
<td>15.72</td>
<td>37.72</td>
<td>41.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>11.33</td>
<td>36.11</td>
<td>39.25</td>
<td></td>
</tr>
<tr>
<td>Computational</td>
<td>1</td>
<td>36.66</td>
<td>58.33</td>
<td>50.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>36.33</td>
<td>60.34</td>
<td>52.48</td>
<td></td>
</tr>
<tr>
<td>Financial</td>
<td>1</td>
<td>35.64</td>
<td>52.30</td>
<td>42.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>33.33</td>
<td>58.12</td>
<td>45.25</td>
<td></td>
</tr>
<tr>
<td>Legal</td>
<td>1</td>
<td>41.66</td>
<td>60.33</td>
<td>53.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>38.35</td>
<td>63.45</td>
<td>48.86</td>
<td></td>
</tr>
<tr>
<td>Text type</td>
<td>Evaluator</td>
<td>System Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Golden</td>
<td>Al-Muterjim</td>
<td>An-Nakel</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Al-Wafi</td>
<td>Al-Arabey</td>
<td>Al-Arabi</td>
<td></td>
</tr>
<tr>
<td>Medical</td>
<td>1</td>
<td>34.21</td>
<td>63.33</td>
<td>43.93</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>35.86</td>
<td>66.14</td>
<td>42.12</td>
<td></td>
</tr>
<tr>
<td>Petroleum</td>
<td>1</td>
<td>40.03</td>
<td>61.28</td>
<td>48.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>41.22</td>
<td>63.42</td>
<td>52.13</td>
<td></td>
</tr>
<tr>
<td>Military</td>
<td>1</td>
<td>43.33</td>
<td>53.33</td>
<td>20.24</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>48.16</td>
<td>50.06</td>
<td>17.87</td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>1</td>
<td>43.75</td>
<td>52.42</td>
<td>36.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>42.09</td>
<td>58.11</td>
<td>38.26</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td>1</td>
<td>60.00</td>
<td>58.33</td>
<td>50.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>65.43</td>
<td>62.22</td>
<td>52.75</td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>1</td>
<td>60.31</td>
<td>57.96</td>
<td>53.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>65.55</td>
<td>60.33</td>
<td>55.36</td>
<td></td>
</tr>
<tr>
<td>Water Engineering</td>
<td>1</td>
<td>33.66</td>
<td>51.66</td>
<td>36.66</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>35.25</td>
<td>56.48</td>
<td>42.02</td>
<td></td>
</tr>
<tr>
<td>Average Value of Eval. 1</td>
<td></td>
<td>39.22</td>
<td>53.61</td>
<td>41.62</td>
<td></td>
</tr>
</tbody>
</table>

**APPENDIX (F)**

The Experimental Values of Texts % Syntax
<table>
<thead>
<tr>
<th>Category</th>
<th>Eval 1</th>
<th>Eval 2</th>
<th>Eval 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical</td>
<td>30.00</td>
<td>32.35</td>
<td>30.17</td>
</tr>
<tr>
<td>Computational</td>
<td>53.00</td>
<td>50.66</td>
<td>48.55</td>
</tr>
<tr>
<td>Financial</td>
<td>38.46</td>
<td>40.35</td>
<td>46.15</td>
</tr>
<tr>
<td>Legal</td>
<td>44.00</td>
<td>42.95</td>
<td>48.00</td>
</tr>
<tr>
<td>Medical</td>
<td>38.18</td>
<td>34.22</td>
<td>44.54</td>
</tr>
<tr>
<td>Petroleum</td>
<td>40.00</td>
<td>38.86</td>
<td>41.37</td>
</tr>
<tr>
<td>Military</td>
<td>32.72</td>
<td>30.43</td>
<td>20.36</td>
</tr>
<tr>
<td>Physical</td>
<td>37.27</td>
<td>39.88</td>
<td>36.36</td>
</tr>
<tr>
<td>Social</td>
<td>50.00</td>
<td>53.86</td>
<td>45.00</td>
</tr>
<tr>
<td>Commercial</td>
<td>43.80</td>
<td>45.73</td>
<td>40.61</td>
</tr>
<tr>
<td>Water Engineering</td>
<td>36.00</td>
<td>38.63</td>
<td>44.00</td>
</tr>
<tr>
<td>Average Value of Eval. 1</td>
<td>39.75</td>
<td>42.75</td>
<td>40.40</td>
</tr>
</tbody>
</table>
APPENDIX (G)

Statistical Aspects

The (Pearsons’ Linear Correlation Coefficient) usually has a symbol $r$ and is a pure number without units or dimensions and it always lies in the interval -1 to +1, the following formula is well adapted for use with the pocket calculator or specific statistical softwares:

$$ r = \frac{n \sum XiYi - (\sum Xi)(\sum Yi)}{\sqrt{[n \sum Xi^2 - (\sum Xi)^2][n \sum Yi^2 - (\sum Yi)^2]}} $$

$n$ = number of individuals

$Xi$ = dependent variable

$Yi$ = independent variable

$r$ is a pure number without units or dimensions and it always lies in the interval -1 to +1

❖ When two variables are inversely related, it means that as one increases, the other decreases.

❖ The absence of any relationship between variables is denoted by a correlation coefficient of 0.00 or thereabouts.

❖ When pairs of values plotted in the way that they fall along a straight line from the lower left of the scatter gram to the upper right, then the correlation coefficient is equal to + 1.00 and when the points fall along a straight line, which runs from the upper left part of the
scatter plot to the lower right, then this indicates a perfect negative relationship, a value of -1.

- The size of \( r \) reflects the amount of variance that can be accounted for by a straight line, whether the data are essentially linear or not.

- Positive values indicate a tendency for both variables to increase (or decrease) together, while negative values indicate that high values of one variable are associated with low values of the other variable.

- The levels of significance most frequently encountered are 0.05, 0.01, and 0.001.
- The level of significance =0.05 means that 5 percent of the values are significant.