College of Arts and Sciences

Department of English and Translation Studies

TRA 699 (Thesis)

A Global Perspective on Machine Translation: Arabic as a Case Study

Submitted in Partial Fulfillment of the Requirements of the Degree of Master of Arts in English/Arabic/English Translation and Interpreting.

By

Wesam AL-Assadi

Supervisor

Dr. Rana Raddawi

Spring 2004
Declaration

I. Wesam...Fawzi ...Al-Assael., confirm that the work carried out for the purpose of this thesis and its text have been done by me. I am the sole responsible for any errors or misrepresentations of facts. I also understand the rules in force at the American University of Sharjah (AUS) pertaining to plagiarism and academic integrity.

Date: May 11th, 2004

Signed: [redacted]
# Table of Contents

Acknowledgements  
Dedication  
List of Abbreviations

## Introduction

1. Translation in the Global World  
2. Machine Translation

1.1 Translation in the Global World  
1.2 Machine Translation

1.2.1 Historical Review  
1.2.2 Defining Machine Translation  
1.2.3 Different Types of Machine Translation  
1.2.4 Types of Machine Translation Demands  
1.2.5 Popular Conceptions about Machine Translation  
1.2.6 Optimizing Machine Translation Efficiency

1.3 Linguistic Strategies in Machine Translation

## Chapter One: Insights into Machine Translation

2. Translation in the Global World  
3. Machine Translation

2.1 Translation in the Global World  
2.2 Machine Translation

2.2.1 Historical Review  
2.2.2 Defining Machine Translation  
2.2.3 Different Types of Machine Translation  
2.2.4 Types of Machine Translation Demands  
2.2.5 Popular Conceptions about Machine Translation  
2.2.6 Optimizing Machine Translation Efficiency

2.3 Linguistic Strategies in Machine Translation

## Chapter Two: Approaches for Machine Translation: Theories & Applications

3. Approaches for Machine Translation: Theories & Applications

3.1 Computational Linguistics and Machine Translation  
3.2 Applied and Theoretical Components of Computational Linguistics  
3.3 Multi-linguality Initial Problem for Theories  
3.4 Language Engineering  
3.5 Linguistics and Computational Complexities of MT

3.5.1 Types of Linguistics Complexities in CL  
3.5.1.1 Specific Words  
3.5.1.2 Morphological Analysis  
3.5.1.3 Syntactic Structures  
3.5.1.4 Semantic Roles and Features  
3.5.1.5 Real World Knowledge  
3.5.1.6 Stylistic Matters

3.6 Machine Translation In Use

3.6.1 A Brief Global Overview  
3.6.2 MT in the United States  
3.6.3 MT in Europe  
3.6.4 MT in Japan  
3.6.5 MT in India
### Chapter Three: Machine Translation in the Arab World

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 The Crisis of Arabic Language</td>
<td>51</td>
</tr>
<tr>
<td>3.2 Complexities of Arabic Processing as a Natural Language</td>
<td>52</td>
</tr>
<tr>
<td>3.3 Theoretical Approaches to Arabic Processing</td>
<td>60</td>
</tr>
<tr>
<td>3.4 Arabic Language Engineering</td>
<td>61</td>
</tr>
<tr>
<td>3.5 Computational Processing of Arabic</td>
<td>62</td>
</tr>
<tr>
<td>3.5.1 Models of Research Projects</td>
<td>63</td>
</tr>
<tr>
<td>3.5.1.1 Finite-State Analysis and Generation of Arabic at Xerox</td>
<td>63</td>
</tr>
<tr>
<td>3.5.1.1 Finite-State Theory and Tools</td>
<td>64</td>
</tr>
<tr>
<td>3.5.1.2 Arabic Finite-State Morphological Analysis</td>
<td>65</td>
</tr>
<tr>
<td>3.5.1.3 Advantages and Availability of Finite-State Implementation</td>
<td>68</td>
</tr>
<tr>
<td>3.5.1.2 Towards Understanding Arabic: Logical Approaches for Semantics</td>
<td>69</td>
</tr>
<tr>
<td>3.5.1.2.1 Arabic Understanding</td>
<td>69</td>
</tr>
<tr>
<td>3.5.1.2.2 Semantic Processing</td>
<td>70</td>
</tr>
<tr>
<td>3.5.1.2.3 Semantic Representation</td>
<td>71</td>
</tr>
<tr>
<td>3.5.1.2.4 Semantic Composition</td>
<td>72</td>
</tr>
<tr>
<td>3.5.1.2.5 Logical Form</td>
<td>73</td>
</tr>
<tr>
<td>3.5.1.2.6 Compositional Rules</td>
<td>76</td>
</tr>
<tr>
<td>3.6 The Automation of Arabic Language: Academia vis-à-vis Industry</td>
<td>79</td>
</tr>
<tr>
<td>3.6.1 Historical Overview</td>
<td>79</td>
</tr>
<tr>
<td>3.6.2 Arab Research Institutes and MT</td>
<td>81</td>
</tr>
<tr>
<td>3.6.3 Active Companies in the Field of MT</td>
<td>87</td>
</tr>
<tr>
<td>3.6.3.1 Sakhr</td>
<td>87</td>
</tr>
<tr>
<td>3.6.3.2 Coltec</td>
<td>89</td>
</tr>
<tr>
<td>3.6.3.3 L&amp;H Appteck</td>
<td>90</td>
</tr>
<tr>
<td>3.6.3.4 Cimos</td>
<td>90</td>
</tr>
<tr>
<td>3.6.3.5 ATA</td>
<td>90</td>
</tr>
<tr>
<td>3.7 List of Commercial Arab Machine Translation Software systems</td>
<td>91</td>
</tr>
<tr>
<td>3.8 Arabic Translation 'Engines' on the Web</td>
<td>91</td>
</tr>
</tbody>
</table>

### Chapter Four: Corpora Analysis

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Theoretical Skeleton</td>
<td>94</td>
</tr>
<tr>
<td>4.2 Points of Interest in MT</td>
<td>98</td>
</tr>
<tr>
<td>4.3 Data Preparation</td>
<td>98</td>
</tr>
<tr>
<td>4.3.1 Language Combination</td>
<td>98</td>
</tr>
<tr>
<td>4.3.2 Text Types</td>
<td>98</td>
</tr>
</tbody>
</table>
4.3.3 Users' Needs
4.3.4 MT systems

4.4 Evaluation Process
4.5 Analysis

4.5.1 First Medical Text
4.5.2 Second Medical Text
4.5.3 Technical Overview / Information Technology
4.5.4 A News Article

4.6 Conclusion
4.7 Recommendations

Conclusion
References
Acknowledgements

Needless to say, many people stood by me in the writing of my thesis. But three names stand out: Dr. Rana Raddawi who was my advisor, and Dr. Basil Hatim and Dr. Said Faiq, who were my professors during my MA program.

To them I owe most thanks and to them this work is dedicated.
Behind this work lied the support of a man, my husband, for patiently bearing with my work demands and odd hours.

To my daughters Farah and Sara, thank you for allowing mum to be with her papers instead of mothering you.

Thanks also to my father for his encouragement and to my mother for praying this work go to a successful conclusion.
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HT</td>
<td>Human Translation</td>
</tr>
<tr>
<td>MT</td>
<td>Machine Translation</td>
</tr>
<tr>
<td>HAMT</td>
<td>Human-Aided Machine Translation</td>
</tr>
<tr>
<td>MAHT</td>
<td>Machine-Aided Human Translation</td>
</tr>
<tr>
<td>FAMT</td>
<td>Fully-Automated Machine Translation</td>
</tr>
<tr>
<td>CL</td>
<td>Controlled Language</td>
</tr>
<tr>
<td>TM</td>
<td>Translation Memory</td>
</tr>
<tr>
<td>CL</td>
<td>Computational Linguistics</td>
</tr>
<tr>
<td>ACL</td>
<td>Applied Computational Linguistics</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural Language Processing</td>
</tr>
<tr>
<td>LE</td>
<td>Language Engineering</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>LT</td>
<td>Language Technology</td>
</tr>
</tbody>
</table>
INTRODUCTION

There is no need to do more than mention the obvious fact that a multiplicity of languages impedes cultural interchange between the peoples of the earth, and is a serious deterrent to international understanding.

(Warren Weaver, 1949)

The twentieth century has been called the 'age of translation' (Jumpelt, cited in Newmark, 1988, p. 3) or 'reproduction' (Benjamin, cited in Newmark, ibid). Whereas in the nineteenth century translation was mainly a "one-way means of communication between prominent men of letters and, to a lesser degree, philosophers and scientists and their educated readers abroad.... Trade was conducted in the language of the dominant nation...Diplomacy was in French" (Newmark, 1988), translation in the twentieth century became a prominent factor in a world moving towards multi-lingualism, global economy and global knowledge.

In today's global economy, as more people increasingly need or want to communicate with counterparts abroad, global inhabitants are confronted with daunting cultural and language divisions. The fact that most of the world's people cannot communicate in English, the predominate language of international commerce, communications and publications, is enforcing a global digital divide.
Globalism has intensified the demand for translation. To reach customers around the globe, businesses must offer information and instruction in the language of the target customer. If the world's non-English speakers want to be more than passive receivers of information, translation is needed. The impact of the Internet has been significant in recent years. The increasing multilinguality of the web constitutes additional challenges for translation industry. The global web can only be mastered with the help of multilingual tools. Today, there are many systems designed specifically for the translation of Web pages and electronic mail. The demand for immediate translations will surely continue to grow rapidly and users are already seeing an accelerating growth of real-time on-line translation on the Internet itself, and all this translates into urgent need for more translation. But there aren't nearly enough translators to cope and the need is urgent for the machine to help.

When computers appeared at the end of the Second World War, there were great hopes of the potential benefits which the imagined powers of these 'electronic brains' might bring. One was the prospect of translating languages to break down communication barriers and to further the cause of international peace.

The early dreams that stimulated research and development efforts was that of a machine that would produce high-quality translation from a wide variety of languages at a low cost. Even the supporters of machine translation agree that decades of effort have not produced the breakthroughs necessary to achieve this dream. Supporters of machine translation say that we would be closer to the dream if we had not given up so soon. Negative evaluations of machine translation in the 60s were based on the argument that the understanding of text by computer
was too difficult. The ALPAC report by the American National Research Council concluded that the basic technology for machine translation had not been developed, and recommended a focus on long-term research in computational linguistics and improvements of translation methods.

One of the original aims of applied computational linguistics was fully automatic translation between human languages. Through bitter experience, scientists have realized that they are still far away from achieving the ambitious goal of translating unrestricted texts. Nevertheless, computational linguists have created software systems that simplify the work of human translators and clearly improve their productivity. The expectation that translation machines might replace people has been replaced with the view that these technologies are instead tools to enhance the efforts of professional translators and researchers. Today the challenges of machine translation improvement illustrate the broader challenges of information technology research, development and use.

Changes in thinking about machine translation reflects the evolution of new concepts of how machine translation systems might be developed and used. Progress in Natural Language Processing Technology, the development of more powerful computers, the increasing availability of large dictionary data sets and advances in some aspects of linguistic theory suggest opportunities for research and development.

Most MT research and virtually all commercial MT activity has concentrated on the major international languages: English, French, German, Spanish, Japanese and Russian. The
languages of the less developed countries have been largely ignored. Yet it can be argued that the need for MT in these countries is as great, or perhaps greater than in the more developed countries.

Arabic, meanwhile has its own dividing line. It can, on the one hand, lead to a wider linguistic divide between the Arabs and the rest of the world at various levels including linguistic studies and language computation or, on the other hand, constitute a pivotal factor getting on board the information train.

Access to sources of knowledge in languages other than Arabic is mainly connected with translation. Translation into Arabic is still extremely scarce and is not keeping pace with the global knowledge explosion. The lag emphasizes the importance of developing machine translation.

Computation of the Arabic language has been hindered for a long time because Arabic systems were designed according to the model of English processing. This model has proven ineffective when used for Arabic for a simple reason: the computation of the Arabic language, compared with English is much more complicated on each level of the language matrix. This has prompted some Arab researchers to design computerized models using the Arabic language as a superset, supplanting English.
This thesis aims to examine the field of machine translation since it was first launched in the fifties and the development it has witnessed since then. It also aims to emphasize the Arab world problems that hinder MT development in Arabic and to propose solutions.

The choice to focus on MT and computational linguistics in the Arab world is for various reasons. First, it is a domain which has not been researched enough both at the undergraduate and postgraduate levels. Therefore, students of translation and the future Arab translators have very little idea about this fast improving field. Second, Arabs have yet to ride the wave of one of the most profound technological phenomena in the history of mankind – the Internet. Despite the ongoing debate among Arabs on the technological revolution brought about by the Internet worldwide, the fact is that the use of Arabic on the Internet is relatively in its embryonic stage still. Even the use of Internet itself is not as widespread in the Arab world as it is in other parts of the world mainly due to language. MT can be a major beneficiary vehicle to over-ride language barriers on the Internet. This will help most sectors of the Arab population to access the Internet which constitutes today the major source of information. Third, as for my self working in the field of journalistic translation, I found that MT is a field that merits examination since most Arab journalists who have no command of English have found themselves forced to use MT systems, both commercially and on-line to translate the content of Internet sites so as to collect information. According to many such journalists, the output is unreadable. However, many claim they have managed to at least get the gist of the information they are looking for.
It is obvious then that MT and language technology are most needed in the Arab world in order for it to engage in the Information Age through translation. Machine translation and localization is a flourishing industry in the West as well as in Japan and other countries around the globe in facing globalization. The objective of this thesis is to see to what extent MT has been developed and utilized in the Arab world.

The thesis is divided into four chapters. The first two are dedicated to machine translation in general, whereas the other two are dedicated to machine translation in the Arab world. The first chapter examines machine translation as a concept, its demands, strategies, types and future expectations. The second chapter examines machine translation as application designed by using theories of Computational Linguistics, Language Technology and Natural Language Processing (language Engineering). Chapter two also covers the status of MT research and systems around the globe. Chapter three is dedicated to the status of machine translation in the Arab world, the crisis of the Arabic language in the Information Age, complexities of the Arabic language in processing within the guidelines of Computational Linguistics and Language Engineering. The chapter also includes a brief overview, which covers the status of MT in the Arab research institutes. It also includes a list of the pioneer Arab and international companies interested in the research and application of Arabic language technology and MT, as well as a list of some of the commercial MT systems available in the market. Chapter four is dedicated to a corpora analysis of texts translated by commercial MT systems from English into Arabic. The aim is to examine the standard of MT output in order to expose the weaknesses and the strength of the Industry and to try to propose solutions. It seems that MT in the Arab world is still in its preliminary stages and a lot of work, time and financial resources are required to further improve the results.
My research in machine translation in the Arab world wasn't conducted without impediments: First, resources are very scarce and the number of books on this topic are but a handful, so my main source of information was the Internet. My main concern was objectivity and accuracy of information. Second, it proved difficult to get responses from Arab research and academic centres with regard to research on machine translation and language technology in the Arab world. Due to the lack of coordination from these institutes, the scope of contact was limited to reliable sources of information.
Chapter One

Insights into Machine Translation

Research into machine translation has already celebrated its fiftieth birthday, yet understanding of its success and failures is still minimal. Even the increase in availability of machine translation software due to the globalization of the Internet has had little impact. The User's knowledge of the complexities behind translating remains limited and judgments are based on personal experience. For more than five decades, people have tried to program computers to translate from one natural language to another. However, since the earliest days of computing, automatic machine translation of natural language has always been an impossible dream, a controversial topic, a source of illusions, jokes and even serious disputes.

This chapter aims to bring forth into the arena some of the crucial issues behind machine translation. Understanding of these particular issues is the only way to move closer to the dreams of a society no longer hindered by language barriers.

Topics related to machine translation, on theoretical and application levels will be covered: translation in the global world, machine translation types and demands, popular conceptions about machine translation and how to optimize machine translation. A brief historical overview is also included.
1.1. Translation in the Global World

It is assumed that translators, more than any other professionals, feel the real changes brought about by the information age. The global market, the increase in intercultural contacts and the acceleration of information production, have resulted in profound changes in the way translators work.

Currently, human translators must use an extensive knowledge base to achieve the main task of translation – the transfer of technical and cultural information. As such, translation requires new strategies and a paradigm shift in methodology. "This shift must embrace practice, teaching and research," argues Austermuhl (2001, p.1).

The concept of globalization in the sense that we – the globe's inhabitants – are citizens of a "global village", entails a debatable question: why bother with more than 4000 different languages if we may do with one language, English? Since English is the dominant language now in business, sciences, technology and international politics, "it is the lingua franca of the global market economy," according to Austermuhl (2001, p.2). Around 85 percent of international organizations use English as their working language. In Europe, 99 percent of international organizations have English as one of their official languages (Mai & Welch, 1999, p.130 cited in Austermuhl, 2001). In addition, around 90 percent of all scientific publications are written in English. Around 98 percent of German physicists publish in English. Even in France, two thirds of scientists use English to publish their research results addressing the global audience (Raethel, p.1, cited in Austermuhl, 2001).
Moreover, language now is a major factor in the debate over globalization, especially because the Internet has made its political, cultural and economic importance universally clear. In 2001, around 80 percent of the contents of the over one billion Internet pages on the web were in English. The 8,000 on-line databases currently available are extracted from information originally published in the English language. “Concern over the future of linguistic diversity in the Information Age is evident from the currency of such terms as ‘language divide’, ‘extinction of languages’, ‘linguistic racism’, and ‘linguistic wars’ (Arab Human Development Report, 2002, p.76)”.

In this context, Austermuhl (2001) raises questions like: “Is English ringing the death knell for the rest of the world’s languages? Will the vision of monolingual world lead to the end of translation? Most probably, the answer is no”.

Politically, the experience of the European Union over the last 50 years supports the view that the need of translation is not new. In Europe multilingualism is a fact of life, each of the 15-member states of the EU is entitled to use its own language to conduct official business within the institutions of the EU. “This institutionalized multilingualism is made possible by the work of 4,000 in-house translators, interpreters and terminologists, and many other free lancers. Each additional official language increases the demand by 250 to 300 linguists” (Stoll, 1999, p.17, cited in Austermuhl, 2001). Since there are 11 official languages and 110 possible language pair combinations, it is not surprising that in 1997, 2 billion euros were spent on translation, both human and machine in the institutions of EU (including interpretation and terminology work) (Austermuhl, 2001, p.3).
Beyond political institutions, in fact, knowledge of foreign languages is not widely spread in Europe. Around 28 percent of German executives have very good command of English skills. A university study conducted in 1999 indicates that one in four German university professors would not attend international conferences if English were the sole working language.

It is relevant to observe here that facility with the English language is waning across the Arab world. "With the exception of a few university professors and educated individuals, real proficiency in English has ebbed, preventing many Arab researchers from publishing their research in international scientific journals", according to the Arab Human Development Report (2003). This trend also explains the wide reluctance to make presentations at scientific gatherings in English, or to participate in seminars or even Internet user groups.

It is obvious then that language diversity vis-a-vis English as a lingua franca of the Information Age actually increases the need for translation.

The increasing cross border communications, the rapid growth of technical and scientific production and the concept of a global market have led to the accelerating growth in the international demand for translation. Austermuhl (2001) cites Germany as a good example for the size of such growth. The German market, he argues has been witnessing a constant 14 percent annual increase in translation for several years. In 2001, the total annual translation demand from German market reached 30 million pages.

The increase in the demand for translation is also partially due to the shift in the Internet from English language only to an international platform for communication and information.
Austermuhl argues that the non-English speakers are "the fastest growing groups of new Internet users, with a rapidly growing interest in non English sites as the Net becomes genuinely multilingual. Websites in Spanish, Portuguese, German, Japanese, Chinese and Scandinavian languages are showing the strongest growth rates" (Austermuhl, p.5).

Although around 57.4 percent of the Internet users were basically English speakers in 1999, there is evidence that the number of the non-English speaking Internet users is rising steadily as penetration rates in non-English speaking countries continue to rise. According to Computer Industry Almanac (cited in www.escawa.org., 2004), the number of Internet users surpassed 530 million in 2001 and will continue to grow strongly in the next few years. Most of the growth coming from Asia, Latin America and parts of Europe. By the year 2005, the number of worldwide Internet users will exceed 1 billion. According to Diab (2003), while the Arabic language population constitutes 18.1% of total world population, the estimated number of Arabic language Internet users is 0.8% of the total world users.
The following figure shows the distribution of online language population totaling 561 million (March 2002)

![Online Language Populations](image)

Figure (1): Distribution of the online language population

This in turn means that the number of multilingual sites will grow and translation services and software are becoming an integral part of international communication. International Data Corporation (cited in www-1.ibm.com, retrieved on Feb 23rd, 2004) estimates that the machine
translation software market sales record were around $378 million in 2003, according to Beck, the IBM Voice Systems Director.

Not only globalization and the increasing numbers of non-English Internet users have caused the growing demand for translation but the digitization of the global economy has a lion's share in this developing industry. Translation is now closely related to the changes going on in the field of international business and communications. These changes are in fact, influenced by the use of modern means of communication and information technologies.

Austermuhl (2001) argues that translation is also influenced by the enormous degree of technical specialization and economic diversification taking place today. He provides selected "factoids" which reflect the size of the information explosion taking place now:

- The amount of knowledge to be processed within the next decade is larger than the amount of knowledge accumulated during the past 2500 years.
- 165,000 scientific journals are currently being published;
- 20,000 scientific papers are produced every day (Mark, 1998, cited in Austermul, 2001);
- The amount of data that is circulating on the Internet on any given day is larger than all the information available throughout the nineteenth century (Der Spiegel, 1996, cited in Austermuhl, 2001)

The previous figures indicate that the size of the information flood is too large for the human brain to process on its own. Humans definitely need the service of electronic tools; the aid of the computer to conduct translation.
1.2. Machine Translation

1.2.1 Historical Review

The idea for a machine that would transfer one language to another came from code breaking during the WW II by in 1949. According to Bass (1999), Cold War intelligence spurred the development of machine translation due to the great amount of documents in Russia gathered by the U.S. military and intelligence agencies during the 50’s and 60’s. By the end of the 60’s the interest in MT began to fade and funding for research stopped until late 70’s.

The American National Academy of Sciences published a report by its Automatic Language Processing Advisory Committee widely known as the (ALPAC) report. The report recommended that research on MT should stop immediately due to its failure to produce useful translation. The ALPAC, report though widely condemned as biased and short sighted, hindered MT research for a decade in the US and in the Soviet Union and Europe as well. However, research continued in Canada, France and Germany. In 1970, MT Systems were installed for use by United States Air Forces (USAF). In the same year, another successful operational system appeared in Canada: the Meteo System for translating weather reports, which was developed at Montreal University.

By then, the advances in theoretical linguistics allowed more sophisticated approaches in MT and resulted in the first practical MT tools for mainframe systems. The impact of the personal computer revolution that began in the 1980’s has opened the ground for the development of PC-based machine translation software (Bass, 1999).
The earliest systems consisted primarily of large bilingual dictionaries where entries for a SL gave one or more equivalents in the TL and few rules of how to follow simple syntax. A number of MT projects were developed in parallel with the developments in the field of linguistics, particularly, in models of formal grammar, according to Hutchins.

The 1980's witnessed the development of a wide variety of MT systems and from a number of countries. In addition to Systran, now functioning in many pairs of languages, there were Logos (German-English and English-French), the METAL system (German-English) and some Japanese-English and English-Japanese systems developed by Japanese companies (Hutchins). The wide availability of microcomputers and of text processing software encouraged the creation of cheaper MT systems, such as: ALPS, Wildner and Globalink. Other systems were developed by Japanese companies such as Sharp, NEC, Mitsubishi and Sanyo.

In the 1990's, MT systems, based purely on statistical method and corpora approach, were developed. In both methods, no syntactic or semantic rules were used in the analysis of texts or in the selection of lexical equivalents. Over the last few years, the use of MT and translation tools has grown tremendously, especially in the era of software localization. There has been also a huge growth in sales of MT software for personal computers (especially among non-translators) and more significantly there has been a major increase in availability of MT from on-line networks. More rapid growth is seen nowadays for direct Internet applications (electronic mail, web pages, etc).
1.2.2 Defining Machine Translation

Machine Translation is the "application of computers to the translation of texts from one natural language into another" (Hutchins, 1986, p1).

According to the European Association for Machine Translation, MT is "the automation of translation" (Napier, 2000). Machine Translation, also known as automatic translation or mechanical translation is "the computerized methods that automate all or part of the process of translating from one language to another," according to Seasly (retrieved on 11th March 2004). It is a multi-disciplinary field of research. It incorporates ideas from linguistics, computer science, artificial intelligence, statistics, mathematics, and many other fields. If machine translation, Seasly argues, becomes accurate and efficient enough, it can break down cultural barriers and make communication between speakers of different languages much easier. Commercially, machine translation can allow companies to translate product manuals more quickly into the target language, or target languages. Thus machine translation can expand a company's Market, save translators' time and companies' money.

1.2.3 Different Types of Machine Translation

Using an appropriate terminology, there are four basic types of machine translation. See the following figure adapted from Hutchins and Somers (1992, p. 148):
The first two types represent the two extremes in translation; human translation is carried out without the help of the machine, and the fully automated machine translation, also called unassisted MT, takes pieces of text and translates them into output for immediate use with no human involvement. The other two types are categorized under assisted MT. The HAMT is the MT that uses human help and the MAHT is where the humans use machine's help.
1.2.4 Types of Machine Translation Demands

One can distinguish four types of machine translation demands and use of computer based translation software, according to Hutchins (1999):

1- The use of MT for **dissemination**. This is the first and traditional type, where the demand is for quality translations as expected from human translators, i.e. translations of publishable quality. However, MT systems still may produce output which must invariably be revised or 'post-edited' by human translators. In this sense, MT systems are mostly producing 'draft' translation.

2- The use of MT for **assimilation**. This type of MT is required for translations of lower level of quality (particularly in style). It is used by users who want to find out the essential content of a particular text, and generally as quickly as possible. The users here feel that they would rather get some translation, no matter how poor. With wide spread cheaper PC based systems, this type of MT has grown rapidly.

3- The use of MT for **interchange**. This is a demand for translation between participants in one-to-one communications (telephone or written correspondence). This type is typically required for translations of electronic texts on the Internet, such as web pages, electronic mail and even electronic ‘chat’ lists. The need here is merely to convey basic content, hence for immediate translation regardless of quality.
The use of MT for information access. This is the integration of translation software into systems for search and retrieval of full texts or documents from database systems for summarizing texts. This field is currently the focus of a number of projects in Europe in order to widen the access of all EU member states to sources of information.

1.2.5 Popular Conceptions about Machine Translation

Austermühl (2001) argues that there is a public perception about MT that swings between two extremes:

1) MT is a total waste of money & time. The quality of output is generally very poor. The traditional anecdote here is that of the Russian MT system that translated *The spirit is willing, but the flesh is weak* into the Russian equivalent *The vodka is good, but the steak is lousy.*

2) MT will break all language barriers in the global stage. In just a few years time the output of machines will be as good as humans’ output.

Although machine translation may not provide a complete solution to the problems of translation due to the unique and complex nature of natural language, it can be an efficient tool in translation of text, at least in a restricted knowledge domain. In addition, MT can help the human translators to improve the speed and productivity of translation. It is certainly unjust to consider MT useless in practice. The professional use of MT requires certain tools to improve the quality, when needed.
In chapter four, some sample texts will be translated by using MT software systems to show how helpful machine translation can be and where it needs improvement.

1.2.6 Optimizing Machine Translation Efficiency

Different approaches can be taken to optimize MT efficiency:
- Human Interaction before (pre-editing), during and/or after (post-editing) MT,
- Controlled Language (CL),
- MT combined with Translation Memory (TM) systems,
- Dictionary building and updating,

A very brief explanation of the approaches follows:
- Pre-editing is the process of identifying problems and, where necessary, editing the ST before translating it, so as any strings of text that an MT system will have problems with, are highlighted and removed or modified in advance.
- Post-editing is a step or a set of steps in an overall translation process for editing, modifying and/or correcting machine-translated texts.
- A Controlled Language, by definition, is a subset of a natural language whose grammar and dictionaries have been restricted to reduce or eliminate ambiguity and complexities in texts written in that CL.
- Translation Memory Systems are basically the building of a translation database for a given document or group of documents. TM software 'records' or stores previously processed texts for display when needed, according to Belis (retrieved on 15th Dec 2003).
Dictionary-building and updating is common place in many translation offices and service bureaus. Many translators, translation offices and companies build their own in-house dictionaries in order to ensure consistent usage. On the other hand, there are now numerous multilingual terminology for specialized scientific, technical, administrative and economic fields, according to Flanagan (2002).

1.3 Linguistic Strategies in Machine Translation

The general strategy employed in nearly all MT systems until the late 1960’s was the direct translation approach: systems were designed in all details specifically for one pair of languages i.e., in most cases, for Russian as SL and English as TL. Hutchins (1986, p.34) argues that:

The basic assumption was that the vocabulary and syntax of SL texts should be analyzed no more than necessary for the resolution of ambiguities, the correct identification of appropriate TL expressions and the specification of TL word order.

Syntactic analysis was almost limited to the recognition of word classes (verbs, nouns, adjectives, etc.) to distinguish homonyms for example, semantic analysis, if included, was restricted to the use of features such as ‘male’, ‘concrete’ ‘liquid’, for resolving collocational ambiguity.

In the ‘second MT generation’, however the direct approach was abandoned and replaced by inter-lingual or transfer approaches where translation is indirect via an intermediary language (inter-lingual) or via a transfer component operating upon “deep syntactic” or semantic representation. Whereas in the direct systems, the analysis of SL text is determined by the requirements of the TL text production, in ‘inter-lingual’ and ‘transfer’ systems, the analysis of SL texts is quite independent of the TL. As Hutchins (ibid) states, “The systems are not
therefore designed for translation only between two specific languages, but can in principle be adopted for translation between other pairs of languages by the addition of new programs of SL analysis and TL synthesis.

The ultimate aim then was to develop 'deep structure' representations embodying what was common to two languages and hence to make the first steps towards 'universal' representations. After the shortcoming appeared in a number of MT systems, MT researchers were convinced that "the approach to linguistic modeling adopted or assumed by the 'pure' linguistic theorists is not appropriate," (Hutchins, 1986, p.12).

In order to achieve practical objectives of producing quick translations of technical documents, it is time to take a more pragmatic stance: use the computer to do only what it can do well (accessing large dictionaries, making morphological analysis and producing simple 'rough' parsing) and to use human skills for the more complex problems of semantic analysis, resolving ambiguities and selecting the appropriate expression from a choice of possible translations. In recent years, there has thus been a number of 'interactive' MT systems under development. Interactive systems are most attractive where there is a need for simultaneous translation of a single SL text to a number of languages; the expensive involvement of a skilled human analyst is then employed to achieve the best results.

All MT systems described so far are essentially syntax-based, with semantic analysis operating after the syntactic structures have been determined. Few systems were able to deal with all cross sentence pro-nominalization and semantic links between sentences - those feature which
make a sequence of sentences into a cohesive whole have been neglected (Haliday and Hassan, 1976)

A subject of current interest in machine translation (in the United States in particular), according to Foster et al, is the rapid development of systems for novel language pairs. At least one of the languages in question is taken to be previously unknown to system developers “who must either acquire the necessary knowledge and technology or devise methods that will mitigate the effects of their absences... this is viewed as a welcome counter to excessive reliance on exceptionally large and ‘clean’ parallel corpora”, argue Foster et al (2003).

Researchers today are competing to improve the quality and accuracy of the translations. Although statistical-based translations are not perfect, especially in regard to grammar, this technology is giving the chance for scientists to crack scores of languages in a fraction of the time, and at fraction of the cost, that the traditional methods involved. Scientists at Johns Hopkins are developing statistic-based machine translations of such languages as Uzbek, Bangali, Nepali and others. The ambition is to develop systems for as many as 100 languages within few years. Although the grammatical structures of languages like Chinese and Arabic make them hard to analyze statistically, it will only be matter of time before such hurdles are overcome.

In order to further develop MT systems, it is essential to see how far the theories of computational linguistics serve the field of MT.
Chapter Two
Approaches to Machine Translation
Theories & Applications

The linguist's contribution to the modeling of translation process has been always recognized in translation studies. What is specifically needed in the development of MT systems is the hybrid efforts of both linguists and computer science experts in what is recently called Computational Linguistics.

2.1 Computational Linguistics & MT

Approaches in machine translation are very diverse as shown in Chapter One. Some researchers see MT as a means of demonstrating their theories or formalism, with their measure of success based on whether or not the system is an accurate model of the human brain or simply an "elegant" theory. The search for a universal grammar for all languages or translation based solely on neural nets to simulate the human mind fall under these theoretical approaches.

In reality, the method used in computational linguistics basically consists of seeking both theory and practice. A great deal of effort is still needed to create functional MT approaches. Research in MT is still, experimental but guided by solid theoretical foundations. Its sole performance criterion is to obtain results for a well-defined need. There is no one solution for all languages. For every need, a fitting solution must be found.
Computational linguistics (CL) is “a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty”, according to Uszkoreit (retrieved on March 2nd, 2004). It belongs to the cognitive sciences and overlaps with the field of Artificial Intelligence (AI), a branch of computer science aiming at “computational models of human cognition”.

CL, according to Thompson (1985), is that part of the science of human language that uses computers to aid observation of, or experiment with, language. If “Theoretical linguists... attempt to characterize the nature of a language or Language or a grammar or Grammar‘, then theoretical Computational Linguistics proper consists in attempting such a characterization computationally’. In other words, CL concentrates “on studying natural languages, just as traditional linguistics does, but using computers as a tool to model (and, sometimes, verify or falsify) fragments of linguistic theories deemed of particular interest” (Boguraev et al, 1995).

2.2 Applied and Theoretical Components of CL

Theoretical CL takes up issues in theoretical linguistics and cognitive science. It deals with “formal theories about the linguistic knowledge that a human needs for generating and understanding language”, according to Uszkoreit (1985). Today these theories have reached a degree of complexity that can only be managed by computers. CL develops formal models simulating aspects of human language features and implements them as computer programs.

In addition to linguistic theories, findings from cognitive psychology play a major role in simulating linguistic competence.
Applied CL focuses on the practical outcome of modeling human language use. The methods, techniques, tools, and applications in this area are often subsumed under the term Language Engineering or (Human) Language Technology.

The goal of CL is to create software products that have some knowledge of human language. They are urgently needed for improving human machine interaction since the main obstacle in the interaction between human and computer is a communication problem. “Today computers do not understand our language but computer languages are difficult to learn and do not correspond to the structure of human thought,” according to Uszkoreit. Even if the languages the machine understands and its domain of discourse are very restricted, the use of human language can increase the acceptance of software and the productivity of its users.

2.3 Multi-linguality: Initial Problem for Theories

Users communicate with the computer in French, English, Arabic, German or another human language. Multilingualism obviously presents problems for any theory that assumes a text to be embedded in a particular language.

The comparative work carried out by nineteenth century grammarians was concerned with establishing an explanatory basis for the relationship between languages and groups of languages primarily in terms of a common ancestor. The comparative grammar, in contrast, is “concerned with a theory in grammar that is postulated to be an innate component of a human brain”, according to Nabil Ali (retrieved on Jan 8th, 2004). In this way, the theory of grammar is
a theory of human language and hence establishes the relationship among all languages, not just those that happen to be related by historical "accident" (for instance, via common ancestry).

One characteristic of modern linguistics has been the attention given to the formalization of descriptions of linguistic systems. While vigour and precision have been a feature of the writings of linguistics since the late nineteenth century (e.g. The Neogrammarians), it is primarily the work of Chomsky (1957, 1965) which has placed universal grammar at the centre of theoretical linguistics, according to Hutchins (2004).

Formalization assumes that language is, at least potentially, a well-defined system - a view which, not all linguists share. Some would argue that they are still uncertain what kind of grammar is appropriate for natural language and what the general characteristics of the formal model should be. In consequence, much of the theorizing in linguistics about the form of grammars and about the formal treatment of particular linguistic phenomena (case relations, semantic features, transformational constraints, pronominalization, passivization) is carried out in a vacuum with no direct contact with real linguistic data.

In general, linguists have tended to ignore problems of translation. "The theory of translation is one of the least developed areas of modern linguistics", Hutchins (1979). The common attitude can probably be summarized as: "we cannot yet describe linguistic process involving one language only, let alone attempt to describe what goes on in translation" he argues. Why then should machine translation be regarded as a suitable test bed for linguistic theory? The principal reason is that whether a text produced by a MT system is a reasonable translation of another text
another language and can be evaluated by independent judges. It provides a clear test of the
effectiveness or wrongness of a proposed system, since the output in a second language can be
assessed by people unfamiliar with the internal formalism and methods employed (Wilks 1975,
cited in Hutchins, 1986). The evaluation of translations has its problems, but in principle it can
be objective, e.g. by observing whether the users of a manual produced by MT can understand
and carry out instructions as well as users of versions of the manual produced by human
translators or by making back-translations of a MT text into the original language and looking at
the differences – a test which can be done by someone knowing only the original language.

According to Hutchins, there are probably many reasons why linguists have generally been
unwilling to be associated with machine translation – ignorance of the ways of the computer,
more interest in theory than in practical work, etc. – “but often it has been from a mistaken
conception of the real aims of machine translation”. The primary stimulus for MT research has
always been the urgent needs for scientists, engineers, technologists, economists, administrators,
etc. to cope with an ever-increasing volume of material in foreign languages.

2.4 Language Engineering/ Language Technology

This section covers the practical or the applied part of CL which includes methods, techniques
and tools of modeling human language by using the computers. In other words, it demonstrates
how human languages are processed in machine as natural languages. Two concepts will be
defined here: natural language processing and language engineering.
NLP is a branch of computer science that studies computer systems for processing natural languages. It includes the development of algorithms for parsing, generation, and acquisition of linguistic knowledge; the investigation of the time and space complexity of such algorithms; the design of computationally useful formal languages (such as grammar and lexicon formalisms) for encoding linguistic knowledge; the investigation of appropriate software architectures for various NLP tasks; and consideration of the types of non-linguistic knowledge that impinge on NLP. It is a fairly abstract area of study and it is not one that makes particular commitments to the study of the human mind, nor indeed does it make particular commitments to producing useful artifacts, according to Uszkoreit (2004).

Language engineering means computation. In early days of language processing, “most, if not all researchers underestimated the complexity of the problem”, according to Uszkoreit (ibid). Many of them tried to find a mathematical characterization of their tasks and solve the problem simply by looking at the input and output of their systems. Most of the early approaches to machine translation fall into this category. These attempts failed very badly. Within years the great majority of researchers became convinced that insights from linguistics – including phonetics and psycholinguistics are needed in order to make progress in modeling the human language user. Traditionally, the main data were collected from invented example sentences, judged and interpreted by introspection.

2.5 Linguistics and Computational Complexities of MT

In the modern world, multi-linguality is a characteristic of a rapidly increasing class of tasks. This fact is most apparent in an increased need for translations and consequent interest in
alternatives'. The main alternatives include partially or fully automatic translation, machine aids for translators and fully or partially automated production of original parallel texts in several languages.

The linguistic and computational complexities of MT are not always apparent to all users or potential purchasers of systems. As a consequence, they are sometimes unable to distinguish between the failings of particular systems and the problems which even the best system would have.

Translation is essentially a problem-solving activity, choices have to be made continually. The assumption in MT systems, whether fully or partially automatic, is that there are sufficiently large areas of natural language and of translation processes that can be formalized for treatment by computer programs, according to Hutchins. Does this mean at the practical level that problems of selection can be resolved by clearly definable procedures? The major task for MT researchers and developers is to determine what information is most effective in particular situations, what kind of information is appropriate in particular circumstances, and whether some data should be given greater weight than others.

In this section, difficulties encountered in MT when translating from one natural language into another will be outlined.
2.5.1 Types of Linguistics and Computational Complexities

There are many challenges to Machine Translation. Some of them are:

(i) the use of other specific words in the same phrase or sentence 
(ii) the use of morphological information 
(iii) the use of information about syntactic functions and relations 
(iv) the use of semantic features and relations 
(v) the use of knowledge about the subject domain 
(vi) the use of stylistic preferences 

2.5.1.1 Specific words

Decisions based on specific words are the easiest to apply and are capable of the highest degree of precision. At the same time, however, there is inflexibility since there is no allowance for inflected variation of forms or for the least variation of word order. Three types of problem will be discussed: compound nouns, idioms, and metaphors.

All translators are familiar with the need to treat compounds as units to be translated. In many cases an attempt to translate each component of a compound noun would lead to ridiculous results: English 'eggplant' is not 'البندوان' in Arabic, but 'بالنخال'. Many potential problems of homonym can be averted by the entry of the relevant words in combination with others in dictionaries.

The word light for example, can modify another noun in at least three different senses: an adjective 'not heavy', an adjective 'not dark' and a noun 'luminescence or illumination'. In theory, every occurrence could have any one of these meanings, but if there are certain words
which regularly occur with it, it would seem perverse not to make use of this fact. Thus many
MT systems include entries for compounds such as a light ship and a light bulb; and indicate
directly the target language equivalent (French ampoule, German Glühbirne). In this way, the
system can avoid a perhaps lengthy disambiguation process to determine which of the two
senses of bulb is intended ('plant bulb' or 'pear-shaped glass') and in combination with which of
the three senses of light; a process which will have to be done every time the compound is
encountered.

Some would argue that the most difficult area for MT must be the apparently unclassifiable
variety of idiomatic expressions. It is a view which has support in the stories of early MT
systems which translated "out of sight out of mind" as 'invisible idiot'.

The perceived difficulty of idioms is that the individual words take on meanings and
connotations, which they do not have, in their literal usages. However, according to Hutchins, it
is precisely because most idioms are relatively fixed expressions consisting of the same words in
the same sequence, that they can be easily translated into comparable idioms – or if none exist
into a literal equivalent.

Like idioms, metaphors can be treated as fixed compound expressions. Among the European
languages, there is a common thread of similar formations, so that even if a metaphorical usage
is not recorded in the dictionary, it may be possible to produce a 'literal' translation which has
the same metaphorical impact. However, it would be a weakness in any MT system if it did not
account easily for many metaphors, which have become standard expressions of the language.
2.5.1.2 Morphological Analysis

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 by the familiar verb and noun paradigms (Arabic, Persian, 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., یاکلن, یاکلون, یاکلن, etc.; and equivalents in other languages).

It should be stressed that any MT system should as a minimum be capable of recognizing morphological forms and of generating them correctly. However, the alignment of equivalences between the verb forms among languages is another matter, particularly when modal forms are involved (must, might, devoir, falloir, mogen, durfen, etc.)

In general, a MT system which cannot go beyond morphological analysis will produce little more than word for word translations. It may cope well with compounds 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.

2.5.1.3 Syntactic structures

The basic structural features are those of dependency and constituency. Examples of dependency are the relations between adjectives and the nouns they modify and between subject nouns and the main verbs of clauses. Any MT system should be able to identify such relations in languages such as French, German and Arabic (Arabic will be discussed later) on the basis of
gender agreement: *les jeunes filles sont venues, die meisten Frauen sind nicht gekommen*. Of course there are complexities in the syntactic analysis. In English, the lack of overt markers of dependency or the ambiguity of those markers which do exist means that greater weight has to be given to the identification of constituency groups, e.g. noun phrases, verb phrases, prepositional clauses and phrases, etc.

Syntactic analysis is based largely on the identification of grammatical categories: nouns, verbs, adjectives. For English, the major problem is the categorical ambiguity of so many words. In essence, the solution is to look for words which are unambiguous as to category and to test all possible syntactic structures. In the case of a sentence such as:

- Prices rose quickly in the market.

Each of the words *prices*, *rose*, and *market* can be either nouns or verbs; however, “quickly” is unambiguously an adverb and “the” unambiguously a definite article and these facts ensure the unambiguous analysis, where *prices* is identified as a subject noun phrase, *in the market* as a prepositional phrase, and *rose quickly* as a verb phrase.

In addition to syntactic problems, this section of semantic roles and features also wraps up difficulties in MT.

### 2.5.1.4 Semantic roles and features

The recognition of implicit relations may well require access to semantic information. It is common to identify two types: semantic roles and semantic features. By the semantic roles in a structure is meant the specific relationships of nominal elements (entities) to verbal elements
(actions or states): a particular noun may be the ‘agent’ of an action, another may be the ‘instrument’ (or means), another may be the ‘recipient’, and another may refer to the ‘location’, and so forth.

Unfortunately, there is no universally agreed set of semantic roles which can be applied without difficulty to any language. Developers of MT systems are usually obliged to draw up their own list. However, the principal difficulty is the identification of roles, Hutchins argues. In English, the main indicators are the propositions, but these can be ambiguous as to the role expressed; *with* can indicate instrument, manner or context:

- the bottle was opened with a corkscrew
- the bottle was opened with difficulty
- the bottle was opened with the meal

*Semantic features* refer to labels such as ‘human’, ‘animate’, ‘liquid’, ‘young’, etc. assigned to lexical elements. They can be used either in conjunction with semantic roles or independently. For example, for the translation of English *eat* into German it might be considered useful to distinguish between ‘human’ agents and ‘non-human’:

- The boy ate the banana → Der Junge hat die Banane gegessen
- The monkey ate the banana → Der Affe hat die Banane gefressen

Such features have to be assigned to all relevant nouns (i.e. all that could be subjects of the verb *eat*); and can be used in other sentences where choices between human and non-human have to be made. As with semantic roles, there is no established set of features which can be applied to
every language. MT developers have compiled their own lists, some are minimal and rigidly controlled and others are extensive or not applied consistently.

2.5.1.5 Real world knowledge

While semantic features and roles combined with syntactic information can go a long way in resolving ambiguities in the source language and in deciding among translation variants, there are numerous instances where what is apparently needed is knowledge about the things and events being referred to. Examples:

(1) old men and women → les vieux et les vieilles
   or: les vieux et les femmes

(2) pregnant women and children → des femmes enceintes et des enfants
   not: des femmes et des enfants enceintes

In (1) we have no idea, out of context, whether "old" applies to both men and women or only to men. But in (2) we do know that "pregnant" cannot apply to children; it is part of our knowledge about women. This knowledge needs to be incorporated in the MT dictionary in someway, probably by limiting the use of "pregnant" to nouns with the semantic features 'female' and 'mature'.

Similar problems arise with relative clauses:
- Peter mentioned the book I sent to Mary

  → mentioned the book (which I sent to Mary)

  → mentioned (to Mary) the book (which I sent)
The first sentence is ambiguous: either the book itself was sent to Mary or the sending of a book to someone else was mentioned to Mary. It is an ambiguity which cannot be solved out of context even in human translation.

We are led therefore to the argument that good quality translation is not possible without understanding the reality behind what is being expressed, i.e. translation goes beyond the familiar linguistic information: morphology, syntax and semantics.

The clear implication, Hutchins argues, is that what is required for the translation of the more intractable problems of analysis and transfer is the availability of a knowledge bank of information which can be referred to during the translation process. It is the approach commonly referred to as that of Artificial Intelligence (AI). For example, given a sentence such

the following occurring in documents relating to computer hardware:

- Remove the tape from the disk drive

The word tape can potentially refer to a ‘magnetic tape’ or an ‘adhesive tape’. An AI-based system would check in its knowledge bank which is most plausible in this context, i.e. it would seek to answer the question whether magnetic tapes can be removed from disk drives, or whether disk drives can contain or have as parts items which are magnetic tapes. If not, then it may check whether ‘adhesive tape’ is plausible, i.e. whether disk drives are things which can be packaged using this item. Clearly, the knowledge bank must contain highly structured information about a wide range of real phenomena, even when documents deal with a quite narrow domain.
The principal reasons for the absence of knowledge banks in MT systems are probably obvious enough. Coverage of any documents other than those within a narrow subject range would clearly require databases of massive proportions. While the computer hardware and the computer software for fast access may well both be already available, the databases are not. These would demand many years of difficult and complex work by many researchers. Therefore, it is not surprising that MT systems are based on well-known techniques of syntactic and semantic analysis and transfer.

2.5.1.6 Stylistic Matters

According to Hutchins, 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 source texts. Of course, human translators can be guilty of this fault as well — although Newmark (1988) considers literalness to be desirable in literary and authoritative texts, as long as the result is in the appropriate style. However, the aim in technical translation is generally to produce texts which read as if they were originally written in the target language. It is quite evident that MT systems do not achieve this goal. Indeed, it can be argued that they should not aim for idiomaticity of this order, if only because recipients of MT output may be led to assume complete accuracy and fidelity in the translation. It does not need stressing that readability and fidelity do not go hand in hand: a readable translation may be inaccurate, and a faithful translation may be difficult to read (Newmark, 1988).

This account has, of course, by no means exhausted all the areas in which MT systems may have difficulties. Since the major problems of MT systems concern ambiguity, homonymy and alternative structures, it has long been recognized that one of the best ways of ensuring good MT
output is to limit the amount of choice in the actual texts submitted to the system or to limit the system itself to specific text types or subject areas. The latter is exemplified by the well-known Meteo system, which was designed for meteorological texts and for nothing else (Chandious 1989). The former is being adopted by an increasing number of MT users, who require texts to conform to certain restrictions of vocabulary and syntax: certain words are to be used in one meaning only, and complex structures are to be avoided.

Hutchins notes that there are well-tested and familiar methods for word recognition, for morphological segmentation and for syntactic analysis. The use of semantic features and roles is also well researched and reliable. With these techniques it is possible to deal with a wide range of linguistic phenomena with reasonable success – but not always without problems. As illustrated, among phenomena which can be relatively easily handled are: idioms and fixed expressions, phrasal verbs, basic word order (both in analysis and in generation), metaphors (when identifiable by specific words), the morphological and the syntactic disambiguation of homonyms, and the resolution of ambiguities by the use of simple semantic features usually spoken. There remain, however, many phenomena of greater difficulty. Some may not occur often in certain text types and some may not be critical for certain users (i.e. they can be handled easily in post-editing or in interactive modes of operation) – how much difficulty they cause depends largely on local circumstances. Among these relatively more difficult phenomena are: prepositions, tense and modality, coordination, subordinate clauses, pronouns, complex sentences, and stylistic variants (both lexical and structural).
Various methods and techniques are currently being developed to improve the output efficiency of machine translation systems. Various countries around the globe are competing to improve MT systems that serve their commercial and political interests.

2.6 Machine Translation in Use

In this section provides a brief overview of MT status in different countries around the globe. It aims to show how MT has served various international users' ends. MT in the United States, Europe, Japan and India are covered. MT status in the Arab world is scanned in detail in chapter three.

2.6.1 A Brief Global overview (United States, Europe, Japan, India)

The surges of interest in machine translation in particular and the various applications of language technology in general around the globe emerges for diverse reasons: on both sides of the Atlantic, multilingualism constitute a major challenge. In Europe, there is a need to address all languages of European citizens, including the language of the new members in the EU from Eastern and Central Europe. In the USA, they feel that they have strong strategic disadvantage: every one understands English, but they do not understand other languages. Therefore they cannot get information from abroad. The USA and Europe have other reasons to embrace MT and Japan and India have their own ambitions in this regard. Theses reasons, ambitions and the MT systems adopted by each country are demonstrated as follow:
2.6.1.1 MT in the United States

It is worth noting that research and development in human language technologies in the United States is taking place within the framework of broader technological initiatives and the large scale of such research serves mainly intelligence and defense programs. Among the most important initiatives are: the High Performance Computing and Communication (HPCC) program (1991–1997) and the Computing, Information and Communication (CIC) Program which started in 1998, with a budget of US$ 1 billion per year, according to Marrai (2004).

Several parties receive support from this budget, amongst which are: the National Science Foundation (NSF), which supports basic research in Speech and Natural Language; the Defense Advanced Research Agency (DARPA), which carries out core technology development; several national agencies including the CIA, FBI, US Air Force, Dept of Energy, National Security Agency (NSA) and others, which develop application.

The area of Speech and Natural Language Processing has been identified as an important sector. Information Technology is handled by the Division of Information and Intelligent Systems, with its own program on Speech and Natural Language Processing, which has a US$ 3 to 4 million budget per year. There are also Inter-Agency programs such as Human-Computer Interaction (Stimulate Program) or Knowledge Distributed Intelligence (KDI).

On the other hand, US businessmen, researchers and product developers and policy makers need a better understanding of what is going in Japan, one of the main competitors in the world's
technological platform. A minute fraction of the American community can speak and read Japanese. Growing recognition of the importance of technical information produced in Japan has stimulated interest in the role MT might play in making it possible for Americans to access reports of new inventions, products and financial developments in Japan.

The United States is ahead of Japan in some areas. For example, the US currently leads Japan in technological diversity, that is the variety of approaches to MT, and linguistic diversity, that is the number of language being developed. Traditionally, the US has been a pioneer in scientific research in NLP, but research funds in the US have been decreasing. Funding in Japan and Europe has been increasing and will surpass the US level, if it has not already done that, according to Carbonell et al (2003).

2.6.1.2 MT in Europe

MT systems in Europe have been much slower than expected; "markets are small and fragmented, and professional translators are hostile", Hutchins states (2003). Machine translation systems are used primarily by large translation services and by multinational companies.

Some of the notable recent installations in multilingual companies to mention are: Ericson, where the Logos system is providing 10% of translation needs for producing manuals and documentation in French, German and Spanish; SAP, using METAL for German-English translation and Logos for English-French (totaling some 8 million words per year); and Siemens
providing a service based on METAL. The European Commission, the use of Systran continues to grow, (now amounting to some 200,000 pages per year).

Commercially, most of the PC-based MT software originates from Japan and the United States, and sales have been lower in Europe. However, there are notable European products: the Compendium and TI systems (Sail Labs), Personal Translator PT (linguatec), the iTranslator series (originally Lernout & Hauspie, now Mendez), the Reverso systems (Softissimo), the range of ProMT systems (for Russian to/from English and German); and the PARS systems for Russian and Ukrainian to and from English. According to Hutchins (2004), most of these systems are available in different versions for large enterprises, for independent professional translators, and for occasional (home) use, e.g. for translating Web pages and emails.

Other PC-based systems from Europe include PeTra for translating between Italian and English; the Al-Nakil system for Arabic, French and English; the Winger system for Danish-English, French-English and English-Spanish; and the TranSmart system for Finnish-English from Kielikone Ltd. (ibid.).

Since Europe has not reached a significant position in the development of MT systems, Japan has surpassed it both in MT research and MT system production.

2.6.1.3 MT in Japan

In Japan, machine translation is viewed as an important strategic technology that plays a key role in Japan’s increasing participation in the world economy. As a result, several of Japan’s
largest industrial companies are developing MT systems, and many are already marketing their systems commercially. There is also an active MT and natural language processing research community at some of the major universities and government/industrial bodies.

"It is no surprise to find that half of the world's MT research is found on that densely populated archipelago (Japan)", states Brace (2004). Japan's appetite for information, its comparative lack of foreign language skills and its distinguished capabilities in the arena of developing electronic products, drive them to develop an ever-competing machine translation systems industry.

As a result, several of Japan's largest industrial companies are developing MT systems. The principal use of MT in Japan is in translating technical documents for products to be sold abroad. While many Japanese MT systems have been developed by protégés of Nago, the systems in practice do vary from the nearly direct Pense, system of Oki to sophisticated, semantically rich systems like Toshiba's Astransac and Fujitsu's Atlas. Other systems boast a wholly different lineage, notably newcomer Logo Vista.

While known for their technical abilities, the Indians do not share with Japan the utmost need for MT since most Indians speak English. Nonetheless, India is entering the Information Age with confidence and MT signifies a vital step towards playing a role in globalization.

2.6.1.4 MT in India

In India, there is a big market for translation between English and the 18 constitutional languages there. Currently, this translation is essentially manual. Use of automation is largely
restricted to word processing. Today the Indian Ministry of Information Technology has realized the importance of MT and has identified specific domains for the development of MT systems, such as government administrative procedures and formats, parliamentary questions and answers, pharmaceutical information and legal terminology and judgments (Srikanth et al., 2001).

"In India's multi-linguistic landscape, where the need to facilitate smooth communication between the Centre and the states is vital for good governance, machine translation offers a great solution to this problem", argues Srikanth (ibid). The social or political importance of MT arises from the socio-political importance of translation in countries where more than one language is spoken. Since most information is in English, machine translation has emerged as a critical technology that can help communication and share information more effectively.

However, machine translation in India is relatively young, according to Raa (2003). The earliest efforts date from the late 80s and early 90s. The most prominent among these are the projects at IIT Kanpur, University of Hyderabad, NCST Mumbai and CDAC Pune. The Technology Development in Indian Languages (TDIL), an initiative of the Department of IT, Ministry of Communications and Information Technology, Government of India, has played an instrumental role by funding these projects. Since the mid to late 90's, a few more projects have been initiated – at IIT Bombay, IIIT Hyderabad, AU-KBC Center Chennai and Jadavpur University Kolkata. There are also a couple of efforts from the private sector – from Super InfoSoft Pvt Ltd, and more recently, the IBM India Research Lab (ibid.).
Another field which has witnessed a rapid improvement is the field of World Wide Web where countries around the globe need to communicate in a diversity of languages; hence on-line translation is rapidly growing.

2.6.1.5 MT on the Internet

The Internet has produced a rapidly growing demand for real-time on-line translation. The need is for fast acquisition of foreign-language information where top quality output is not essential. Many PC-based systems are marketed for the translation of Web pages and of electronic mail, and there is great and increasing usage of MT services (often free), such as the well-known ‘Babelfish’ on AltaVista. At the same time, the Internet is providing the means for more rapid delivery of quality translations to individuals and small companies. A number of MT system vendors currently offer translation services, usually ‘adding value’ by human post-editing.

Finally, the Internet has also demonstrated an urgent need to replace the existing systems, developed for well-written scientific and technical documents and assuming human post-editing, by systems and translation aids which are developed specifically to deal with the kind of colloquial (often ill formed and badly spelled) messages found in emails and chat rooms, where there is no possibility of any human revision. “The old linguistics rule-based approaches are probably not equal to the task on their own, and we may expect corpus-based methods making use of the voluminous data available on the Internet itself to form the basis of future systems for this application”, argues Hutchins (2004).
In short, today there are several MT systems in different forms available for various languages. These MT systems differ in their functional structure and the methodology of formulation taking into consideration the nature and complexities of languages involved in the process. These MT systems acquire significant practical importance due to the explosive growth and usage of the Internet in the areas of on-line business research, education, communication and in the government. Some of these MT systems provide a faster and cheaper translation in addition assisting human translators improving their productivity and efficiency in translation.

Arab countries are no exception in this regard. Arabic is the mother tongue of over 300 million people in 22 Arab states. If the Arab world is to be a knowledge based society in which all its organizations and all of its population can participate, it is essential then to develop websites which can be accessed in Arabic. MT is the magic tool for several reasons: Since the international websites are overloaded with information written in hundreds of languages, only machines can translate millions of words daily. Access to the Internet is essential for the economic development of Arab countries. Commercial MT systems will help in the acceleration of translating technical and scientific books for which there is a demand in the Arab world.

The next two chapters are dedicated to MT in the Arab world, its research approaches and its applications.
Chapter Three

Machine Translation in the Arab World

Access to sources of knowledge in languages other than Arabic is mainly connected with translation. Translation into Arabic is still extremely scarce and is not keeping pace with the global knowledge explosion.

According to the Arab Human Development Report (2002) issued by the United Nations Development Program (UNDP), Arab countries annually translate around 330 books, which constitute one-fifth the amount of books translated in Greece. The accumulated number of books translated since the ninth century is around hundred thousand books. This number equals what Spain translates in one year.

This attitude to translation is in direct contrast to the status of translation in the Medieval Arab World. At that time, translation, according to Faiq (2000), played a vital role in the establishment of Arab-Islamic cultural and intellectual identity. It "made the Arabic language a world linguistic medium of knowledge for many centuries".

It is possible to compare the Arab present time with the medieval era in terms of the need to adopt knowledge and sciences from foreign civilizations. Medieval Arabs recognized the importance of translation for their endeavors to strengthen their new state, and translation then became a matter of official concern. Arabs today are in critical need at assimilating knowledge and of building a systematic Pan-Arab translation programs to meet the information explosion of
this era in history. Because Medieval Arab translators were under pressure, they adopted three main strategies of translation; transliteration, literal and gist translation. Each strategy was used according to the specific needs of the time. Transliteration was used in the very beginning of the translation movement then, literal translation was used in order to gain as much information as possible in a short time and gist translation was used at the point when the need for more translation diminished when Arab scholars started writing and publishing their own research. Translators then worked with linguists and grammarians to coin Arabic equivalent terms. On the other hand, Medieval Arab translation flourished in the eighth century when Arabs began producing paper on a large scale.

Today, Arabs have the opportunity to use the electronic tools and media (as compared to paper in the Medieval Age), to help in the assimilation of knowledge in no time. Using machine translation is essential if the Arabs is to compete in this globalized world. If the outcome of MT translation is unacceptable (again transliteration, literal and gist translations are some of the strategies used in MT), it is always possible to improve the outcome with human aid. There is an utmost need today for translators, linguists and grammarians to unify their efforts in building advanced MT systems.

It is obvious then that Arabs are not any more in a phase of time to debate whether we need to use MT systems or not, but rather to improve MT programs to better serve their needs, Raddawi (2004) argues.

This chapter covers issues of interest regarding Arabic in machine translation. Also included in the chapter are: the crisis of Arabic language, computational processing of Arabic, theoretical
approaches to Arabic, Arabic as a Natural Language and language engineering and research programs proposing solutions to the complexities of Arabic as a Natural Language. A survey is given to the Arab research institutes interested in machine translation; research and applications and pioneering companies in the field and finally a list is provided for the Arabic commercial machine translation software systems.

3.1 The Crisis of Arabic Language

Language is today a recurring topic in the debate over globalization, especially now that the Internet has made its political, cultural and economic importance universally clear.

Linguistically, the world of information and communication technology is at a watershed. It can maintain linguistic diversity, a choice that entails difficult communication and hinders flow of information and knowledge, or it can turn to a standard unified language, most probably English.

Arabic, meanwhile, has its own watershed. This language can become a means for Arab countries to catch up with the information train, or it can lead to a wider linguistic divide between the Arabs and the rest of the world at various levels, including linguistic studies, lexicography, language education, the professional use of language, the documentation of language and language computation.
Arabic today, on the threshold of a new knowledge society, faces severe challenges and a real crisis in terms of theorization, teaching, grammar, lexicography, usage, etc. The rise of information technology presents a real challenge to the Arabic language today.

According to the Human Development Report 2003, issued by the United Nations Development Program (UNDP), central to the Arabic language crisis are the following: first, there is a marked absence of linguistic policy at the national levels, which diminishes the authority of language centers, limits their resources and eventually results in poor co-ordination among them. Second, the Arabicization of the sciences and various other disciplines has not proceeded according to expectations. Third, there is a chronic deficiency in translation efforts in the sciences and the humanities. Fourth, Arabic linguistic theory suffers from stagnation, isolation from modern philosophical schools and methodologies, and a lack of awareness of the role language plays in modern society. Fifth, the situation of Arabic language is further complicated by the duality of standard and colloquial Arabic. Sixth, Arabic electronic publication is weakened by the scarcity of advanced Arab software. Finally, the Arabic language continues to suffer from the duplication of research and development projects and the absence of co-ordination among them, “conflicting diagnoses of the ills afflicting the language, and the conspicuous absence of a clear vision of linguistic reform” (HDR, p.123).

3.2 Complexities of Arabic Processing as a Natural Language

Arabic, as a Semitic language, differs from European languages morphologically, syntactically and semantically. There has been much interest recently in the handling of morphologically rich inflectional languages such as Arabic from a computational perspective.
Several workshops in recent years (both regional and affiliated with international conference) have addressed the spectrum of issues relating to the processing of Arabic. The progress over the years has opened the door to advanced computational applications such as machine translation. Research of machine translation of Semitic languages is still, however, in its early stages. Accurate translation of Arabic and other Semitic languages requires treatment of unique linguistic characteristics, some of which are common to all Semitic languages; others are specific to each of these individual languages.

Natural Language Processing is needed because around 75% of all information is textual. In order to process information computationally, we need first to process texts computationally. In 1983, according to Ali (2004), Arabic was extremely unprivileged in the computation field, "suffering the limitations of a minimal system at a pure character level and poor printing and display qualities. Thus it was necessary to shift to a more developed level dealing with larger linguistic units, namely the word, the sentence, and the continuous text". As an expert in the field of MT in the Arab world, Ali said that Arab researchers followed the steps of English as the most established computation example, because "we had to draw on its resources and techniques". Shortly after starting their research, Arab researchers discovered that these techniques were not suitable for Arabic. This is simply due to the fact, according to Ali, that Arabic as compared to English is "much more complex at almost all linguistic levels, with phonology as the sole exception" (ibid).
Figure (3): Prepared by Ali (ESCWA, 2003) to demonstrate Arabic Assimilation within English.

Since the objective of research into natural language processing is to make computers deal "intelligently with the diversity and complexity and variation of human natural languages", according to Yaseen, et al (2003), Arabic natural language processing is considered one of the most difficult among the Semitic and non-Semitic languages due to the complexity of its automatic processing. Research in Arabic NLP is very rich in areas such as morphology, moderate in syntax analysis and still not very mature in semantics and lexicon building.
For the last two decades concentration on Arabic language processing has focused on the manipulation and processing of the structure of the language from morphology and syntax point of views. According to Yaseen, et al (ibid), achieving Arabic understanding requires more than that. In order to achieve natural language understanding a differentiated and deep semantic processing is required.

Chalabi, Head of Sakhr Research Centre in Egypt, told the researcher via telephone and e-mail that since the Arabic language is computationally one order of magnitude more complex than its Latin counterparts, it is unrealistic to import solutions developed to process less complex languages like English and French so as to adapt them to handle Arabic. On the contrary, Chalabi said that Sakhr, after developing its own Arabic NLP components which took more than 15 years, with an average team of fifty linguists, engineers and designers, decided to adopt the same components to process English. While it took Sakhr 2 years to develop a full-fledged morphological analyzer for Arabic, only 3 months was needed to develop the corresponding morphological analyzer for English.

According to Chalabi, some of the major problems in Arabic NLP are:

- **On the character level:**
  - a) Character context sensitivity
  - b) Overlapping
  - c) Diacritics and points

- **On the word level:**
  - a) Highly inflectional language
  - b) Different writings for some characters (Alef, Maksoora and Hamza).
On the syntax level:

a) lack of diacritics in written text.

b) Free word order

c) Rare use of punctuation

Chalabi argued that native solutions specifically built to tackle the Arabic language have proven to be more efficient, reliable and most of all more salable than their counterparts borrowed from English. However, not even Chalabi claimed to have solved all the problems. Some of the problems still needing research, according to Chalabi, are:

- Part of speech disambiguation
- Word sense disambiguation
- Pronominal reference solution
- Elliptic personal pronouns detection
- Named entity detection

For Ali, (1994, p.355), the complexity of Arabic at the character level lies in the cursive shape and concatenation of Arabic letters, and above all these letters are characterized by a high degree of context sensitivity. By this, it is meant that its appropriate shape is determined by the surrounding letters (note the changing shape of the “Ain” according to its place: ) (ع، ع، ع، ع، ) At the word level, the morphology of Arabic is “the most sophisticated of all languages”, according to Ali (1994, p.354). Complexity in the Arabic morphology becomes very clear in its acute derivational aspect. Lastly at the syntactical level, Arabic has no doubt proved to be the most difficult, primarily because Arabic is usually written without vowels. Arabic syntax is also
recognized for its wide syntactic transformation mechanisms like anaphora and cataphora ex:

بَلَغَ القَطْمَ لُنَا رَضِيعًا instead of بلَغَ لُنَا رَضِيعَ القَطْمَ

substitution and ellipsis (such as using the subject noun instead of the verb) ex: يَضْبَرُ الْوَلَدُ أَخاَهُ ضَمَارِبًا أَخَاهُ instead of ضَمَارِبًا أَخَاهُ. According to Ali, in order to process English syntax computationally, around a thousand arithmetical rules were used, whereas more than twelve thousand rules were used for syntactic Arabic computational processing. Ali argues that, in essence, written Arabic is “a quasi-stenographic script, and this results in a severe melange of various ambiguities, which are unprecedented and absent from any other languages”, argues Ali (1994). The morphological ambiguity is due to absence of vowels is intermixed with other types of ambiguities, mainly those associated with word sense, part of speech and syntactical structure. Ali provides an example to explain such a problem:

Assumed sentence: 'some firms lend money'.

The sentence as would be written in the Arabic fashion:

"SM FRMS LND MNY" (Ali, 1994).

The result as it appears is a string of constants, each consonantal forms may have a set of alternative vowelized interpretations. Thus, according to Ali, any syntactical processor dealing with Arabic text as its input has to primarily disambiguate such quasi-stenographic script. As a result, an automatic vowelizer became mandatory as prerequisite for Arabic computation. To solve this problem, Ali has developed an order to disambiguate the unvowelized text, as well as to substitute the missing vowels. This required the achievement of the three main computational linguistic tasks: 1) the development of an Arabic parser; 2) the development of a lexical-semantic processor and 3) the development of an automatic generator of the vowelized text.
Since parsing techniques developed for English have been proven inadequate for the Arabic language, both in function and performance, a parsing system based on a multi level grammar was developed and implemented, according to Ali. "This system is capable of handling the previously mentioned intermixed set of ambiguities. The disambiguation mechanism works incrementally at every level of the grammar. Resident ambiguities are resolved heuristically, resorting to preferential principles working on both syntactic and semantic levels" he argues.
In order to solve the linguistic complexities of Arabic language, especially in regard to its computational processing as a NLP, the next part of the chapter will shed light on various theoretical approaches which serve as a basis to find suitable solutions.
3.3 Theoretical Approaches to Arabic Processing

A distinction can be made between two approaches in machine processing of Arabic. One set of approaches can be qualified as 'particularist' because they "emphasize the linguistic idiosyncrasies of Arabic and use them for a local processing approach. This approach is considered more in agreement with the internal requirement of the Arabic linguistic system", according to Guidere (retrieved on the 7th March, 2004). On the other hand, the 'universalist' approach highlights the actual or assumed possibilities of application of methods already tested for other languages, such as English and French into Arabic, with or without adaptation. Guidere argues that the 'particularist' approaches are concerned mainly with the morphological and semantic aspects of the Arabic language, while the 'universalist' approaches emphasize the syntactic aspects of the linguistic system.

However, Hannach, a specialist in the field of Computational Linguistics, told the researcher in a private interview that unless Arab linguists develop linguistic theories that can cope with the mathematical algorithms of the computer, it will be difficult to develop applications of language technology in Arabic that are efficient and feasible.

3.4 Arabic Language Engineering

Most mainstream language engineering techniques have been developed for Western European languages. These techniques, though superficially quite distinct, are built according to formal algorithms that machine can 'realize'.
In an interview with Al-Khaleej Daily on 12 January (2004), Hannach called for a renaissance in the field of Arabic linguistics. According to him, a challenging task facing the research community in the Arab world is to develop computer algorithms and their applications that can process Arabic texts. Unless a linguistic theory is developed according to the metrics of the new machine technologies, the launching of efficient Arabic automatic applications will remain lagging behind, Hannach argues.

Since computers are essentially logically programmed systems built on strict mathematical algorithms; linguistic rules must be strict and formal, according to Hannach. Computer engineers in the Arab world, having ignored the linguistic side of information and communication technology, they will naturally come up with programs that are unable to compete in the international market and which fail to meet the requirements and expectations of Arab users.

Although MT of Arabic is more difficult in general, according to Hannach, it enjoys some linguistic features that make its automatic processing a task with few complexities. It is basically built on specific roots and patterns for verb forms and for nouns and adjectives derived from verbs. Roots constitute the basic skeleton of words in Arabic, whereas patterns constitute their overall structure. According to Hannach (2004), this mathematical architecture of the Arabic language makes it more 'fusional', in contrast to some other languages which are 'affixational'. 
In order to develop promising Arabic information and communication technology applications (machine translation among them), there is a pressing need to improve the machine processing of Arabic as a natural language.

3.5 Computational Processing of Arabic

Translating between languages as different as Arabic and English is complex, for humans as well as machines. The best translations are not simple word for word translation substitutions, but go beyond the surface structure and transmit the deep meaning and concepts into other languages. Implementing this “knowledge based” translation process requires tremendous effort in programming the computer with the knowledge it needs to translate correctly.

In order to provide users of specific language with easier access to the knowledge, we need to apply natural language processing to the information they seek in their native language. According to Yaseen et al, (retrieved on 29th of December 2003), the objective behind Arabic language processing as a NLP is to provide Arabic speakers access to the “huge Latin accumulation knowledge over the web and across the Internet in Arabic”.

Research activities, both linguistic and technical, are crucial for the development of any machine translation system. In the following part, extracts from two research projects will be provided as examples of linguistics research conducted to improve the shortcomings in machine translation.

3.5.1 Models of Research Projects

This section includes two research studies: Finite-State Morphological Analysis and Generation of Arabic at Xerox conducted at Research Department in Xerox and Towards Understanding
Arabic: *Logical Approach for Semantics* conducted by Haddad and Yaseen. The two papers aim at demonstrating possible solutions to overcome the complexities facing the automatic processing of the Arabic language as a natural language. Each research paper covers different aspects of complexities; the first one covers the morpho-syntactic complexities whereas the second paper examines the semantic problems.

### 3.5.1.1 Finite-State Morphological Analysis and Generation of Arabic at Xerox

Xerox Research Centre in Europe has developed a morphological analyzer based on the Finite-State Technology. A phonological analyzer has been developed to analyze orthographical words that may include full, partial or no diacritics. If diacritics are present, they automatically constrain the ambiguity of the output and a fully vowelled spellings are returned with each analysis. Beesley, K., Xerox Research Centre Europe, described the morphological analyzer in its simplest form as a “black box module that accepts words and outputs morphological analyses” (2001).

In computer analysis of Arabic, or of any other language, the input words are in digital form, with the characters in standard encodings like ASMO8859-6 and Unicode. As for the content of the morphological analyses, they will always be somewhat theoretical – and application-dependent. In short, a morphological analyzer should separate and identify the component morphemes of the input word, labeling them somehow with sufficient information to be useful to the task at hand.


3.5.1.1.1 Finite-State Theory and Tools

A language is a set of strings (sometimes called sentences) made up by concatenating together symbols (characters or words) drawn from a finite alphabet or vocabulary. If a language has only a finite number of sentences, then a complete characterization of the set can be given simply by presenting a finite list of all sentences, according to Kaplan, R., Xerox, Palo Alto Research Center. But if the language contains infinite number of sentences (as almost all languages do), the same sort of "recursive or iterative description must be provided to characterize the sentences".

According to Beesely (2001), lexicons and morphotactic information are encoded in the lexc language, which is a kind of right-recursive phrase-structure grammar, and are compiled into finite-state transducers. Finite-State Transducers (FSTs) are data structures that encode regular relations, which are mapping between two regular languages. For human convenience, a finite-state relation is visualized as having 'upper-side' regular language and a 'lower-side' regular language; and each string in one language is related to one or more strings in the other language, Beesely explains.

The upper-side or analysis strings of an FST compiled from the lexc description consists of underlying morphemes and multi-character-symbol tags like: +Noun, +Verb, +Adjective, +Conjunction, +VPref (verbal prefix), +Masc[cule], +Fem[inine], +Sig[ular], +Plur[lar], etc. that identify the morphemes. These tags have multi-character print names that are chosen and spelled according to the taste and needs of the developers, but they are manipulated internally exactly like the other types of characters. The related lower-side language consists of surface
strings. They may still represent underlying strings requiring the application of alternation rules to map them into properly spelled surface strings, according to Beesely (2001).

Xerox Research Centers conducted studies to apply this technology on Arabic as a means of Arabic morphological analysis.

3.5.1.1.2 Arabic Finite-State Morphological Analysis

In computer analysis of Arabic or of any other language, the input words are in digital form, with the characters in standard encodings like ASM08859-6 and Unicode: As for the content of the morphological analysis, they will always be "theory-application dependent". According to Beesely (2001), the morphological analyzer in the broadest terms, should separate and identify the component morphemes of the input word and label them somehow with sufficient information to be used in the tasks at hand.

As for Arabic, it is presumed that a morphological analyzer would separate and identify prefixed word like morphemes such as the conjunctions (wa-) and (fa-), prefixed prepositions such as (bi) and (li-), the definite article, verbal prefixes and suffixes, nominal case suffixes and enclitic direct-object and possessive pronoun suffixes.

The Arabic morphological analyzer is built using finite state compilers and algorithms, and the results are stored and run as finite-state transducers. The finite-state approach to morphology, using a variety of software implementations, has become very popular around the world, having been used to create morphological analyzers for all the commercially important European
languages, including Hungarian and Finnish, as well as Japanese, Korean, Swahili, Aymara, Malay, etc.

At Xerox, the treatment of Arabic starts with a lexc grammar where prefixes and suffixes concatenate to stems in the usual way, and where stems also are represented as a concatenation of a root and a pattern as shown in Figure (6):

Upper: \([kb\&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg\)
Lower: \([kb\&CaCaC]\)

Upper: \([bny\&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg\)
Lower: \([bny\&CaCaC]\)

Upper: \([qw\&CaCuC]+Verb+FormI+Perf+Act+a+3P+Mas+Sg\)
Lower: \([qw\&CaCuC]\)

Figure 6: Three pairs of strings in the lexicon FST were compiled from the lexc description. These examples correspond to the words that will eventually be \(katabat\) (كتبت) \((banat\) (بنت) and \(qaala\) (قال). The upper-side analysis strings contain the roots \(kb\), \(bny\) and \(qw\) respectively, the verbal form perfect active pattern \(CaCaC\) or \(CaCuC\) and the third-person feminine singular suffix \(at\) or the third-person masculine singular suffix \(a\). The square brackets are used for convenience to delimit the stem components from the rest of the word, and the ampersand serves here as just a delimiter between the root and the pattern, which are simply concatenated together. In other upper-side strings, the various morphemes are separated and are identified with multi-character tags and bracketing conventions. The lower-side strings, still abstract here, will be mapped via finite-state algorithms and alternation rules into properly spelled surface strings.

The first step in the modification of such strings, according to Beesley, is to interdigitate the roots and patterns to form stems, but only on the lower side of the relation. The interdigitation is formalized in finite-state terms as intersection, but it in fact represents a special case of intersection that is performed much more efficiently by a finite-state algorithm called \textsc{MERGE}. 
The application of the merge algorithm to the lower side of the relation is performed by the
COMPILE-REPLACE algorithm and the result is shown in Figure (7):

Upper: [ktb&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg
Lower: katab

Upper: [bny&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg
Lower: banay

Upper: [qwl&CaCaC]+Verb+FormI+Perf+Act+a+3P+Masc+Sg
Lower: qawul

Figure 7: Pairs of strings from the lexicon FST after application of the compile-replace algorithm to the lower side. The lower-side strings, ignoring gaps or epsilons, are now katabat, which is essentially finished (كتابت), and banayat and qawula, which involve weak radicals and await the application of alternation rules to map them into their final orthographical forms (بنات) and (قال) respectively. Note that the upper-side strings have not been modified.

Once COMPILE-REPLACE has been performed on the lower side, the necessary alternation rules can be compiled and applied, via composition, in the usual way shown in Figure (8):

Figure 8: Creation of a Lexical Transducer. The .o. represents the composition operation

Not surprisingly, to anyone who has studied Arabic, the rules controlling the realization of w, y and the hamza (the glottal stop) are particularly complicated. In the examples shown here, katabat is finished and can be displayed as (كتابت), the underlying final y radical of banayat disappears on the surface, leaving banat (بنات), and the underlying medial radical w of qawula disappears as well, leaving qaala (قال), with a lengthened vowel. The state of the string pairs, after composition of the alternation rules, is shown in Figure (9). Further composition of
"relaxation" rules allowing the optional deletion of short vowels and the diacritics completes the picture. The final transducer will directly map from katabat (کتابت) or ktbt (کتبت) or any partially vowelled variation of the spelling to the upper side string [ktb&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg. In the web demo, the various morphemes and tags in the analysis string are separated and reformatted in the HTML for more perspicuous display to the user. The following is figure (9):

Upper: [ktb&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg
Lower: katabat

Upper: [bny&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg
Lower: banat

Upper: [qwl&CaCuC]+Verb+FormI+Perf+Act+a+3P+masc+Sg
Lower: qaala

Figure 9: Pairs of strings from the lexicon FST after composition of the alternation rules on the lower side. The lower-side strings are here displayed contiguously.

3.5.1.1.3 Advantages and Availability of Finite-State Implementations

By keeping within the finite-state domain, grammatical components can be defined, combined and modified using standard finite-state operations. Lexical transducers can be run forwards to generate or backwards to analyze, and they are computationally very efficient for natural-language problems. Xerox Finite-State Morphological Analyzers, running on modern PC and workstations, typically analyze thousands of words per second, according to Beesley (2001). The runtime code that applies lexical transducers to input strings is also completely language-independent. Thus the code that runs the Arabic morphological analyzer is exactly the same code that runs German, French, Spanish, Portuguese, etc.
Xerox’s implementation of Finite-State Theory has been used extensively in its own research and commercial work, and these software tools have been licensed to over 70 universities and non-commercial research centers.

The second study conducted by Haddad and Yaseen examines another aspect of Arabic computational analysis; semantics.

3.5.1.2 Towards Understanding Arabic: Logical Approach for Semantics

For the last two decades concentration on Arabic language processing has focused on the manipulation and processing of the structure of the language from morphology and syntax point of views. According to Haddad and Yaseen (2003), these aspects are very important in the NLP. However, achieving Arabic understanding “requires actually a differentiated and deep semantic processing”. Their project Towards Understanding Arabic: A Logical Approach For Semantic Representation is directed to build a framework for processing the Arabic language in order to achieve the understanding of the language electronically.

3.5.1.2.1 Arabic Understanding

Artificial Intelligence has a long time ago recognized the necessity of performing some semantic inferences. Haddad and Yaseen argue that semantic reasoning based on logical models for Arabic has so far received little attention. There are many morphological analyzers which proved successful in solving morphology related issues. Some success has also been achieved in regard to syntactic issues.
One of the main factors, for this negligence according to the researchers, might reside in the
“complexity of this field and in the invisible collaboration between Artificial Intelligence, Arabists, logisticians and linguists”, according to Haddad and Yaseen. Therefore, it is believed that there is a need to develop an adequate model for understanding and particularly for the semantic processing for Arabic. In spite of the fact that so far no existing formal theory of semantics is able to provide a complete and consistent account of all the phenomena of Arabic, Haddad and Yaseen believe that it is important to develop a model for semantic processing “even if that model is imperfect and incomplete”.

3.5.1.2.2 Semantic Processing

In semantic processing, different basic tasks have to be performed at different levels. These tasks imply: semantic composition (construction of semantic representation for capturing the semantic potential of Arabic propositions), semantic resolution (determining the current semantic value and the disambiguation under using context knowledge) and semantic evaluation (extracting of relevant information based on performing some deductions and inferences on the semantic representation of a proposition under using episodic, rule knowledge and world knowledge).

Furthermore, an interpretation process might need some conceptual knowledge and some pragmatic contents to supplement the meaning of a natural language propositions in a specific domain. For example, concepts like “دراسة” (study), "مشروع" (project) and "فائدة" (interest) need some conceptual knowledge and some pragmatic annotations about their mode and time.
It is, however, important to emphasize that the selected meaning representation formalism plays a central role for the whole semantic reasoning process, according to Haddad and Yaseen.

### 3.5.1.2.3 Semantic Representation

There are many reasons to choose a logical language as a target language for the meaning representation. For Haddad & Yassen, logic represents in particular a well-known meaning representation formalism that differentiates between syntax and semantics. In addition, it enables inferences over quantified descriptions, which are basic requirements for an adequate meaning representation for any natural language.

Furthermore, representing Arabic sentences as logic programs have the unique facility of performing some semantic reasoning tasks on a code based on Arabic predicates. Therefore, embodying logical formulas with Arabic predicates is a very interesting aspect of logic programming in the context of understanding Arabic.

For example formulas like 2.1 compared with 2.2 offer more flexibility in performing some semantic tasks in Arabic sentences.

\[
(2.1) \quad (\exists x) (\text{طالب}(x) \land \exists y (\text{بررس}(x, y)))
\]

\[
(2.2) \quad (\exists x) (\text{تلميذ}(x) \land \exists y (\text{يدرس}(x, y)))
\]

\[
(2.3) \quad (\exists x) (\text{طالب}(x) \land \exists y (\text{درس}(x, y)))
\]
In general, there are some important methodical principles and constraints for any semantic formalism designed for the practical applications. These include Compactness, Modularity, Generality, Expressive Power, Efficiency, Implementation Independence, Theory Independence.

Since Arabic is based on verb-noun and noun-noun opposition, we can establish a correspondence between Arabic sentences and predicate logic first order (PL1) formulas. The verb or the "خيار" of an Arabic sentence (the nominal predicate of a nominal sentence) can be assigned to a predicate argument-structure of the corresponding PL1 formula. The noun phrases can be expressed by constants or by quantified arguments of some predicates representing the role of the subject or the object.

3.5.1.2.4 Semantic Composition

Haddad and Yaseen argue that "a semantic formalism has to be compositional on the level of semantic representation in order to assure the constraint of modularity". Despite the fact that predicate logic corresponds to a well studied and an well-understood formal representation formalism, it does not provide any compositional methods. Based on the type theory of Montague, A-calculus offers a standard framework for filling this gap.

In spite of the importance of Montague’s logical methods in the computational community, these methods are rather constructed to deal with the semantics of sentences and are in general, inappropriate for treating semantic processing of texts and dialogues. One of the most important methods for capturing such problems involved in text anaphoric represents the Discourse Representation Theory (DRT). Combination of DRT with λ-Calculus leads to a compositional framework that is able to capture such problems.
Our current view for achieving natural language understanding in the context of the semantic representation of Arabic, according to Haddad and Yaseen (ibid), is to employ $\lambda$-Abstraction for constructing logical formulas acting as meaning representation for Arabic sentences. $\lambda$-DRT semantic construction will be the next goal for achieving Arabic text understanding.

### 3.5.1.2.5 The Logical Form

Since Arabic distinguishes between different types of sentences: Verbal Sentences (VS), Nominal Sentences (NS) and Copulative Sentences (CS) application of $\lambda$-Calculus requires a contextual interpretation of the meaning of the determiners in the different types of sentences. Because of the fact that the nominal and copulative sentences start with noun phrases, their semantic construction and representation would generally be similar to those sentences in English or German. Semantic composition of verbal sentences requires reordering the compositional process for verbs, with which a verbal sentence normally starts.

#### Noun Phrases

Some difficulties were encountered in capturing the information expressed by the determiners and numerals in noun phrases, Haddad and Yaseen argue. The most used Arabic determiner " $\text{Jill}$ " can be understood as a quantifier. Based on the standard analysis for determiners in the type theory, we can interpret the determiner det (1 $\text{Jill}$, num: sing) as

$$1\text{Jill} \Rightarrow \lambda P \lambda Q \exists x (\forall y (P(y \Rightarrow x = y) \land Q(x)))$$

where $\text{Jill}$ denotes the meaning of the determiner " $\text{Jill}$ ".

The indications of indefinite articles can be interpreted as $\exists$-quantifiers as follows
(3.2) **Indefinite indication** \[\Rightarrow: \lambda P \lambda Q \exists x (P(x) \land Q(x))\]

"\(\exists\)" expresses that there exists only one thing of being P and Q, which implies that the *cardinality* of P has to be 1.

In general, a quantifier differentiates between two things: a *restriction* \(\mathcal{R}\) and a *scope* \(\mathcal{S}\). \(P(x)\) represents in 3.1 the restricted sent (*Restriction or the Base*) and \(Q(x)\) the scope (*the proposition about the restricted set*).

Generally a determiner can be expressed

(3.3) as **Det** \(\lambda R \lambda S (\text{Quantifier}(R,S))\)

In addition, interpreting the meaning of a quantifier requires some *conceptual knowledge* about the relationship between a restriction and its scope and their cardinalities.

For example, the quantifier "\(\exists\)" (most) expresses that the \(|R \land S|\) holds relatively a large portion of \(|R|\).

By treating the "\(\exists\)-quantifier" and numeric quantifiers, we have adopted a similar concept presented in (Binot, 1991, cited in the research paper) by introducing the new quantifiers (: 1\(\exists\)) expressing singular definite determiner, \((1\exists\: )\) expressing plural definite determiner and \((\exists n:\:)\) representing numeric definite determiner.

Determiners like "\(\exists\)" and "\(\exists\)" can be interpreted as all-qualifiers as follows:

(3.4) \[\exists\ \Rightarrow: \lambda P \lambda Q \forall (x, P \to Q)\]

3.4 can also be expressed based on 3.3 as follows:
(3.5) $\lambda p \lambda q ( (R, S))$

"بعض" can be interpreted as $\exists$-quantifier:

(3.6) $\| بعض \| \Rightarrow: \lambda R \lambda S \exists x (x, R \land S)$

**Adverbs** (ظروف) modify verbs and adjectives and therefore they are intentional like quantifiers:

(3.7) $\| \text{Adverb} \| \Rightarrow: \lambda p \lambda q(\text{ظروف}(P, Q))$

Nouns and adjectives in nominal sentences are considered as basic words. They can be represented generally as follows:

(3.8) $\| \text{Noun} \| \Rightarrow: \lambda x \text{ اسم}(x)$

A noun means in 3.8 that there is something, which can have the property of being اسم (noun).

For example applying the meaning of the noun طالب (student), $\| طالب \| \text{ to the proper name }\text{ أيمن} (Ayman)$ means that there is somebody whose name is "Ayman" with the property of being a student:

(3.9) $\| طالب (أيمن) \| \Rightarrow: \lambda x \text{ طالب}(x) (أيمن)$

$\Rightarrow: \text{student (Ayman)}$

Adjectives can be represented similarly:

(3.10) $\| \text{Adjective} \| \Rightarrow: \lambda x \text{ صفة}(x)$

**Verbs**

Verbs in Arabic can be intransitive or transitive. We can represent their meaning as follows:

(3.11) $\| \text{Intransitive Verb} \| \Rightarrow: \lambda x \text{ فعل1}(x)$

(3.12) $\| \text{Transitive Verb} \| \Rightarrow: \lambda x \lambda y \text{ فعل2}(x, y)$
3.5.1.2.6 Compositional Rules

In order to be able to compose logical formulas for Arabic sentences we need to give meaning to structured syntactical categories, like Verbal Sentences (VS) and Nominal Sentences (NS).

It is important to emphasize that in the early stages of performing semantic analysis additional syntactical and semantic information has to be evaluated within the following compositional rules. It is assumed that this information has been obtained by a parser, which will accept only one correct sentence based on the semantic information collected in the lexicon.

The meaning of NS can be obtained by applying the meaning of the "يميّن" to the meaning $||H||$. That means applying of

$||M|| (||H||)$

So if the "يميّن" (M) consists of a determiner and a noun, as it is the case in the following incomplete Logic Grammar, then $||M||$ means the application of the meaning of the noun to $||\text{Det}||$. The meaning of the entire nominal sentence can then be achieved by determining the meaning of "خير", $||\text{خير}||$ and its application to $||M||$.

\[
<\text{NS}> \rightarrow <\text{M} \times \text{H}> \quad \text{sem} \quad ||\text{NS}|| = ||\text{M}|| (||\text{H}||)
\]

\[
<\text{M}> \rightarrow <\text{Det} \times \text{N}>
\]

\[
\text{sem} \quad ||\text{M}|| = ||\text{Det}|| (||\text{noun}||)
\]

\[
<\text{Det} \rightarrow \text{بعض} / \text{كل} / \text{ال} / \ldots
\]

\[
\text{Sem} \quad ||\text{ال}|| = \lambda \text{R} \lambda \text{S} \text{ (x,R} \rightarrow \text{S)}||
\]

\[
||\text{كل}|| = \lambda \text{R} \lambda \text{S} \text{ (x,R} \rightarrow \text{S)}||
\]

\[
||\text{بعض}|| = \ldots
\]

(3.14) \[<\text{H} \rightarrow <\text{Noun}>| <\text{Adj}>| \ldots\]
\[
\text{Sem} \ ||\text{Noun} \ || = \lambda x \ \text{Noun}(x) \\
|| \text{Adj} || = \lambda x \ \text{Adj}(x)
\]

For example the meaning of the noun ميكا/"M (Det (ال, sing), noun (طفل) is the application of the meaning of the noun to the meaning of the determiner

\[
(3.15) \ ||\text{الطفل} \ || =: \lambda R \lambda S ( (x, R^S)) (||\text{الطفل}||) \\
=: \lambda S (\text{ال} (x, \text{طفل}(x)^S))
\]

Applying the meaning of the adjective "جميل" (nice), which takes the role of "خير" yields the meaning of the sentence "الطفل جميل"

\[
(3.16) \ ||\text{الطفل جميل} \ || =: \lambda S (\text{ال} (x, \text{طفل}(x)^S)) (||\text{جميل}||) \\
\lambda S (\text{ال} (x, R^S)) (||||) \\
=: \text{ال} (x, \text{طفل}(x)^S) (||\text{جميل}(x)||)
\]

Considering determiners as quantifiers requires the application of their meanings to the meaning of other syntactical categories. Since verbal sentences start with verbs, and if the "فاعل" (subject) contains a determiner, the meaning of the subject can be achieved by applying the meaning of the noun in the subject to the meaning of the determiner.

In addition, the verb and the object can take the role of the scope of the determiner of the subject:

\[
(3.17) \ <\text{VS}>\rightarrow<\text{VerbXSubXObj}> \\
\text{Sem} \ ||\text{VS} \ || = ||\text{Sub}|| (||\text{Obj}3|||\text{Ver}b||)
\]

||\text{VS}|| i.e. the meaning of VS, is the application of ||\text{Verb}|| to the meaning of the Object and eventually to ||\text{Sub}||.

For example the meaning of "يدرس الطالب الحاسوب" (the student) in the VS "يدرس الطالب الحاسوب" is

\[
(3.18) \ \ ||\text{الطالب} \ || =:
\]
\[ \lambda S(\text{Gold}(x, \text{Student}(x) \land S)) (||\text{Obj}|| (||\text{Verb}||)) \]

Applying of the ||\text{Obj}|| to the meaning of the verb yields:

\[(3.19) \quad \Rightarrow \quad ||\text{Gold}(y,\text{Student}(y) \land S)) \]

Regarding of (3.20) as the meaning of the scope of the determiner in (3.10) yields:

\[(3.20) \quad ||\text{Gold}(x, \text{Student}(x) \land S) (y, \text{Computer}(y) \land S)) \]

\[(3.21) \quad \text{Gold}(x, \text{Student}(x) \land S) (y, \text{Computer}(y) \land S) (x, y))\]

This research paper demonstrates that: First representing Arabic sentences as logic programs has the unique facility of performing some semantic reasoning tasks on a code based on Arabic predicates. Second, achieving natural language understanding in the context of the semantic representation for Arabic is possible through the utilization of \(\lambda\)-calculus for constructing logical formulas acting as meaning representation for Arabic sentences. Third, extending this approach to \(\lambda\)-DRT leads to a good strategy for solving problems involved in text anaphoric and a modular composition, according to Haddad and Yaseen.

It is concluded that the Arabic language exposes certain linguistic complexities for the developers of language processing systems on different levels; syntactic, morphological and semantics. However, what is required is to further research in the fields of Arabic linguistics and language engineering.
In the following section, the contribution of Arab universities and research institutes with regard to research and development of issues related to language technology and computational linguistics will be examined, the activities of Arab industry in this regard will be covered and some of the commercial machine translation software systems available in the market will be listed.

3.6 The Automation of Arabic Language: Academia vis-a-vis Industry

3.6.1 Historical overview

A brief history will be provided for Arabic language automation:

1) In 1962 the National Institute of Planning in Egypt was the first Arab Institute to have a computer (14 years since the first computer was used). As for Arabicization, this computer was used for very primitive functions: to type names and addresses in Arabic and to use Arabic letters to substitute the Latin letters (Ali, 1988).

2) In 1973, a significant step forward was achieved when Said Hayder, a professor at the Montreal University (originally from Pakistan) designed a computational system for automatic recognition of Arabic letters. A system was developed to recover the complexities related to Arabic letters recognition which enjoy high degree of context sensitivity. As a result, the number of Arabic letters on the keyboard were minimized to include the main alphabets’ shapes only such as (ع، ن, ل, ك).

3) From 1973-1985, some important achievements were made:

- The Arabic language was used in the database and information retrieval systems.
- Software systems were developed in Arabic, such as Basic and Logos.
- Preliminary systems for the computational generation of Arabic language were developed.

- The development of partial systems for morphological analysis.

- After ten years of discussions, the Unified Arab Code for electronic data exchange on the Internet received Arab unanimous agreement.

- In mid 1985, the computational processing of Arabic language as a NLP witnessed a turning point on the word level, when Sakhr, succeeded in developing the first software engines or tools for multi-mode morphological and syntactic analysis, diacritization and segmentation.

- The automation of Arabic dictionaries.

- The development of text analysis software which was used in the morphological analysis of the Qur'aan.

- The development of spell-checking systems, the basic tool for word processing systems.

- The development of advanced memory systems where Arab words are stored in their morphologically analyzed shape i.e. using the root and the morphological patterns of Arabic words.

- The development of electronic tools for information automatic retrieval of Arabic. These tools facilitate the search inside Arabic texts for words as they appear in the text without looking up their roots.

- The development of multi-mode syntactic analyzers. Sakhr could develop extensive word lists and a body of 20,000 rules for Arabic grammar and syntax.

In the last ten years, the internationalization of the www and the proliferation of communication tools in Arabic, as shown earlier, demonstrate the need for a large number of Arabic NLP applications. As a result, more research activities have been launched to address
more general areas of Arabic language processing, including syntactic analysis, machine translation, document indexing, information retrieval, etc.

Research in Arabic speech processing has made significant progress due to “more improved signal processing technologies, and to recent advances in the knowledge of the prosodic and segmental characteristics of Arabic and the acoustic modeling of Arab schemes”, Osborn (2004) states. These results should make it possible to further progress in more innovative areas, such as Arabic speech recognition and synthesis, speech translation and automatic identification of a speaker and his/her geographic identification, etc.

3.6.2 Arab Research Institutes & MT

In a telephone conversation with Taher Labib, director of Pan-Arab Centre of Translation in Beirut, Labib told the researcher that machine translation in the Arab world is still a field to be revealed even for most Arab intellectual elites. In the Arab world, debate over machine translation still concentrates on the ability of the machine to translate. According to Labib, a lot of time and effort are still needed to convince the Arab academia, decision makers and the commercial sector of the advantages in using machine translation in the Information Age. According to Labib, there is no consistent and/or systematic machine translation research in the Arab countries. There are individual programs even within the borders of one Arab country, and it is even hard to scan such programs. The Pan-Arab Translation center does not have a record for any machine translation programs or applications available in the Arab countries, according to Labib. The Pan-Arab Centre is preparing a plan for a machine translation program, but it is still in its preliminary stages.
As part of the present study, the researcher tried to contact various research institutes in a number of Arab countries, which are specialized in Information Technology Research and a couple of Arabic language academies to see whether they are working in projects related to research and application of Language Technology. The researcher checked whether or not the academies are working on developing or updating linguistic theories to cope with the requirements of the Global and Information Age, but to the researcher's disappointment, no response was received from any institute or academy.

In the local market, the researcher contacted a number of companies working in the field of Science and Technology the Dubai Internet City to see what kind of research projects they are developing in context with machine translation or language technology and engineering. It was realized that almost all the companies in DIC are basically working in sales and marketing, whereas development and programming are taking place in other countries like Egypt and Jordan. In Egypt, Sakhr was very cooperative. Their research Centre, headed by Chalabi Ashraf, was willing to provide the researcher with the required information.

In the Emirates again, the researcher contacted some research centers specialized in information and communication technologies, but none of them showed an interest in the topic; some of the intellectual elite still believe that the machine is "stupid" and cannot translate and others do not realize the feasibility of using such technology.
Dr. Sultan Al Qasimi, the Ruler of the Emirate of Sharjah, has underscored the urgent need to improve the status of translation in the Arab world. Ambitious translation projects have been launched here in Sharjah in coordination with the Higher Colleges of Technology. However, these projects are confined to the domains of human translation. AUS is encouraging such activity. Prof. Raddawi is heading a project to be launched in Fall (2004) on machine translation and interpretation.

The researcher also contacted the Dubai e-Government to check if they use machine translation in translating their training and public programs. The Information Technology Department told the reporter that they receive their programs already translated.

However, the Centre of Arab Unity Studies mentioned few research activities in the field of language technology and machine translation in some Arab countries in a book entitled *Translation In The Arab World: Towards Establishing a Pan-Arab Translation Centre*.

These research activities are listed in the following section.
The Institute of Electronic Research: The National Council for Research in Cairo

The Institute is executing a program for specialized machine translation in coordination with the European Union to translate medical texts. The program is called ‘ARAMED’. It follows the transfer technique. It is part of the European CATz program. The Institute of Electronic Research in Cairo is developing the Arabic part of the project.

The Institute is also building a multi-lingual dictionary based on CORPUS.

The Institute for Electronic and Computational Research (The King Abdul - Aziz City for Science and Technology)

The Institute was established in 1992 to launch research programs on system Engineering, computational engineering, computer sciences and other related fields.

Some of the research activities conducted there in the domain of the computational processing of Arabic language are:

- The establishment of a database for Arabic texts.
- The development of morphological analyzer for Arabic words.
- The development of automatic diacritizer.
- The establishment of a database for Arabic calligraphy.
Lebanese University/The National Council for Scientific Research:

The researcher Anis Abu Farah, from Lebanese University – now a member of the National Council for Scientific Research has developed a software program for machine translation for Arabic and French. But, for unknown reasons, this program was not published.

Syrian Scientific Research Centre for Information Technology

Fuad Khouri, a member of the Syrian Scientific Research Centre told the researcher via fax that in Syria, machine translation research is still in its very preliminary stages. There is a plan to establish a centre for translation and language processing affiliated with Damascus University, according to Khouri.

However, some research programs in Syria have succeeded in developing assisting tools for language automation, such as:

- A software for Arabic letter recognition. Two systems were developed:
  1) A system which works on the 'VAX - II' and 'IBM-PC',
  2) A system which works on the compatible personal computers ‘IBM-PC’.

The Centre of Arabicization Studies and Research / University of King Mohammed V (Morocco)

The Centre was established to develop Arabicization programs on all levels. Among the other fields of interest, the Institute of the Arabicization Studies and Research has established a department for Machine Translation and Computational Processing of Arabic Language.
In Tunisia, the Regional Institute for the Media and Remote Communication Sciences has developed a Machine Translation System ‘Turjuman’ which will be launched soon.

It is clear that research and development activities in the field of Arabic Language Technology and the Automatic Processing of Arabic language applications are still very few and simple. Arab universities and research centres hardly show any interest in this flourishing field. In order to develop products that will revolutionize machine translation and Arabic language computation software technology, money, time and expertise should be dedicated to integrate efforts exerted by industry to achieve improvements in this context.

Raddawi (2004) stressed the Arab world’s urgent need for a team work where expertise from the fields of translation, linguistics, computer science, engineering and economies work together in order to improve advanced machine translation systems and other applications of language technology.

In addition, Arabs need to build extensive database banks. In order to do so, encyclopedia, Arabic literature, recent books and newspapers and magazines must be scanned to collect idioms, expressions, structures and other features which will enrich our systems, argues Raddawi. Since coordination between the academies of Arabic language is at its minimal level, MT can play an important role in the standardization of terminology among Arab countries, according to Raddawi (2002). "MT systems and software ... and contribute in the process of
standardization of Arabic technical terminology. Consistency can be reached through MT software if put on line and used by everyone” (Raddawi, forthcoming).

Examples:

<table>
<thead>
<tr>
<th>User name</th>
<th>اسم المستخدم</th>
</tr>
</thead>
<tbody>
<tr>
<td>Password</td>
<td>كلمة المرور</td>
</tr>
<tr>
<td>Outbox</td>
<td>البريد الم寇دار</td>
</tr>
<tr>
<td>Folder</td>
<td>ملف</td>
</tr>
<tr>
<td>Toolbar</td>
<td>شريط الأدوات</td>
</tr>
</tbody>
</table>

The Arab academia in general has not realized this need, except for individual efforts. But some Arab and international companies have realized this and have been exerting tremendous efforts to serve the Arab needs in this context.

3.6.3 Active Companies in the Field of Machine Translation

A brief outline will be given for efforts of Arabic and international companies to develop software products of machine translation from Arabic to English and from English to Arabic.

3.6.3.1 Sakhr

Sakhr (a part of Al-Alamiah Group) has developed schemes for machine translation from Arabic-English and English-Arabic.

Over the last 20 years, Sakhr has realized the importance of Arabic Natural Processing as a starting point for Language Technology application. In that foundational approach, Sakhr
developed teams to write formal grammars and to compile lexicons and corpuses of sentences for developing and testing software ‘engines’ to handle Arabic texts, according to Chalabi (retrieved on 29/12/2003). These have provided bases for products as diverse as religious instruction (Arabic versions of the Holy Qur'an, Hadith databases and Arabic tutorials), Internet front-ends, optical character recognition for scanning Arabic text-to-speech applications and machine translation.

For NLP, Sakhr has developed software tools for morphological and syntactic analysis, diacritization and segmentation, plus extensive data sets of words, sentences and grammatical rules. Sakhr has also developed a series of data sets including lexicons based on monolingual (Arabic and English) and bilingual (Arabic-English-Arabic) dictionaries.

The Sakhr machine translation engine is mainly based on the transfer model. Due to the complexity of Arabic language automatic processing, the analysis module, (which is the heart of the MT component) was developed first to handle Arabic then the same techniques have been successfully applied to handle English. Machine translation engine performance has been boosted by a statistical language model contributing in the lexical and morphological disambiguation of the source language, in addition to enhanced word selection on the target language, according to Chalabi. The Sakhr language statistical model is supported by two balanced corpuses one for English and another for Arabic 200 million words.

In an effort to globalize Arabic software industry, Sakhr has insured that their software is compatible with Microsoft Windows and Arabic versions of Windows, which are now the
dominating operating systems for personal computers in the Arab world. In order to open the
door to the Arabic user to efficiently use the Internet, Sakhr has developed a number of tools
and products will be listed later.

3.6.3.2 Coltec

Established in 1990, Coltec is one of the leading companies in the field of Arabic computational
linguistic Research. According to Coltec (2004), the company’s distinguished achievement was
the establishment of a new theory of Arabic language processing that would take into
consideration the linguist systems of non-European languages.

Coltec offers a wide range of solutions for both companies and end users. The Cairo main
branch of Coltec developed the spell checker and grammar checker used for the first time in
1997 by Microsoft word. Coltec has also developed a grammar checker for Microsoft word
2000, tools for word identification, a linguistic model based on statistical techniques. Coltec has
also used heuristic and Artificial Intelligence techniques to build the ‘Hidden Markov Models
(HMM)’ (to extract the Arabic linguistic features required for Information Technology
applications, according to Al-Sabah (2003).

Coltec has also developed morphological analyzer for word and sentence disambiguation, tools
for multi-lingual electronic lexicons, an index for Arabic texts and a system for text retrieval.
3.6.3.3 L & H Appteck

L&H Appteck is one of the pioneering companies in the field of Natural Language Processing worldwide. When Appteck decided to enter the Arab market, the decision was to start strong, so it purchased one of the specialized companies in the field of machine translation research and development: Coltec. The joint company's name became L&H Appteck. In 1990, the new company developed Transphere software for translation from English to Arabic. The software was first developed in 1996 to translate from Arabic into English. The program is based on the Lexical-Functional- Grammar model developed by foreign linguists in the mid-eighties. The software has been further developed and a new version has appeared in multilingual mode and has a translation memory.

3.6.3.4 Cimos

Cimos is one of the leading French companies, which works on the development of machine translation and which considers the Arabic market one of its crucial commercial markets. Cimos' main interest has always been the development of translation and Arabicization services to be installed by other interested companies. Cimos has developed a number of machine translation software systems such as An-Nakel Al-Arabi, Al-Kafi, Al-Mu'utarif, etc.

3.6.3.5 ATA

ATA is a pioneer company in the field of machine translation especially for the Arabic language. The company is based in London. It has developed a number of machine translation software under the well-known commercial name 'Al-Waf'. Its first software was 'Al-Mutafim Al Arabey for professional translation. The company has recently developed a translation
'engine' which uses Artificial Intelligence to solve the linguistic problems in translation, according to Al Marzouki (2002), the Director of Al-Marzouk For Technology and Information, the representative of ATA in Riyadh.

3.7 List of Commercial Arabic Machine Translation Software systems

The following is a list of some of the machine translation software available in the Emirates market working either from English-Arabic, Arabic-English or English-Arabic-English:

- **Al-Mutarjim Al-Arabey**: English – Arabic
- **Golden Al-Wafi v2.00**: English – Arabic
- **Al-Wafi V4.00**: English Arabic
- **Al-Mu'utarif**: English – Arabic – English
- **An-Nakel**: English – Arabic – English
- **An-Nakel**: One way Arabic – English
- **An-Nakel**: One way English – Arabic
- **An-Nakel Multilingual Translation system**.

3.8 Arabic Translation 'engines' on the Web:

The widely used Arabic Translation Web Portals are:

- **Tarjim**, the Arabicization tool on Ajeeb.com,
- **Al-Misbar** developed by ATA
- **CAT Translator**: Bi-directional English – Arabic – English - Sakhr.
- **On-line Translation** - Sakhr.
Since companies like Sakhr, Coltec, Cimos and others claim that they have developed their own linguistic and technical research to develop machine translation systems, it is quite important to examine the output of such products in order to monitor the strong and weak points for future improvement.

In the next chapter a corpora analysis will be conducted on texts selected and translated by two commercial software systems: Al-Wafi (developed by ATA) and Al-Kafi (developed by Cimos). The output analysis will demonstrate the strong sides in the translation of the two systems and the sides which need further research and development.
CHAPTER FOUR

Corpora Analysis

The evaluation of machine translation output has played a crucial role in the development of MT systems since emerging over five decades ago. Although research in machine translation lacks an appropriate, consistent and easy to use criterion for evaluating the results, evaluation tools are indispensable in that they allow us to compare two translation systems or to elicit information as to how a variation of any system might affect the quality of translation. Evaluation of MT system is required, both by developers, before and after system modifications, and by end-users who wish to compare different systems before making a purchase.

The quality of MT translation systems has currently being measured by using a variety of techniques and generally depends upon the context in which the MT system is being used. Whereas many other parameters are relevant to the quality of the system, it is often the output quality that developers as well as users concentrate on.

Organization of this chapter is as follows:

4.1 A theoretical sketch is provided which covers the linguistic guidelines in translating as proposed by pioneer linguists. The aim here is to shed light on MT as a translating process and how it complies to these guidelines. The ultimate goal is to examine and analyze the prepared corpora accordingly.

4.2 General points of reference in MT

4.3 Data preparation.
4.1 Theoretical Sketch: Linguistics & Translating: Human Vs. Machine

In this section a number of linguistic and translation observations are provided in brief and shall serve as a theoretical skeleton upon which the strengths and weaknesses of MT output will be analyzed.

a) According to Hatim and Mason (1990), one obvious application of linguistics is "the attempt to develop a device for carrying out automatic translation" (1990, p.22). The search for fully automatic high quality translation might be expected to provide a point of contact between linguists and the translating profession, "in reality it has largely been a case of separate development" (Hatim & Mason). Instead of initiating a thorough investigation into the actual process as carried out by human translators, early research into machine translation chose to "concentrate on problems of syntactic parsing and resolving lexical politely in sample sentences".

b) An unstated underlying assumption was that translation involved overcoming the contrasts between language systems, source-language syntactic structures had to be exchanged for TL.
structures; lexical items from each language had to be matched and the nearest equivalents selected.

According to Hatim & Mason (1990, p.22)

While a huge investment was made (in terms of both effort and funding) in research into how to resolve such problems, the whole notion of context was deemed to be intractable and, consequently, beyond the bounds of machine processing.

Earlier developments in linguistic theory were of relatively little interest to translators. "Structural linguistics sought to describe language as a system of interdependent elements and characterize the behaviour of individual items and categories on the basis of their distribution" (Hatim & Mason, 1990, p.25). Morphology and syntax constituted the main areas of analysis.

Since meaning is the heart of the translator's work, it follows that the postponement of semantic investigation was bound to create a gap between linguistic and translation studies. "Quite simply, linguists and translators were not talking about the same thing", argue Hatim & Mason (ibid).

Over years, structural theorists, like Catford (1965), attempted to build a theory that emphasizes contextual meaning and the social context of situation in which language activity takes place. However, such attempts are very recent in MT and have not achieved much. In Arabic computation, debate is still ongoing regarding syntax and morphology. This point of interest will be investigated in the corpora analysis.

c) Chomsky's generative – transformational model analyzes sentences into a series of related levels governed by rules. The key features of this model can be summarized as follows:
1) Phrase-structure rules generate an underlying or deep structure which is

2) transformed by transformational rules relating one underlying structure to another to
produce.

3) a final surface structure, which itself is subject to phonological and morphemic rules.

The structure relations described in this model are held by Chomsky to be a universal
feature of human language. Chomsky’s model, as was mentioned before, is the basis upon
which computational linguistics built.

d) Nida and Taber (1969, p.39) claim that all languages have between six and a dozen basic
kernel structures (the most basic sentence structures). Kernels are the level at which the
message is transferred into receptor language before being transformed into the surface
structures in three stages: “Literal transfer”, Minimal transfer, and ‘Literary transfer”
(Munday, 2001). This categorization of transfer will be checked in the corpus analysis since
most MT software systems use transfer as a main MT strategy in translating.

e) According to Nida the “message has to be tailored to the receptor’s linguistic needs...”
(1964, p.159) This is the basis upon which MT evaluation is based. Since MT is used by
various users for various reasons, it is then basically user oriented.

f) Again, according to Nida, the receptor-oriented approach considers adopting grammar,
lexicon and cultural references as essential in achieving “Naturalness”. Naturalness, which is a
‘key requirement’.
For Nida, the success of translation depends on 'four basic requirements of a translation:

1) Making sense;
2) Conveying the spirit and manner of the original.
3) Having a natural and easy form of expressions.
4) Producing a similar response.

g) Literal, or word for word translation is “the direct transfer of a SL text into a grammatically and idiomatically appropriate TL text in which the translators task is limited to observing adherence to the linguistic servitude’s of the TL” (Yinay and Darbelnet, 1958, p.86).

h) According to Yinay and Darbelnet, unacceptable message in translation when translated literally means:

1- gives another meaning, or
2- gives no meaning, or
3- structurally impossible, or
4- does not have a corresponding expression within the metalinguistic experience of the TL, or
5- has a corresponding expression, but not within the same register.

4.2. Points of Reference in MT

General points of interest to MT will be provided in brief to serve the corpora analysis:

- MT is used by different users for different needs. Users' needs usually determine MT output;
whether the user needs a good-polished translation or just a gist translation.
MT is successful in technical and scientific texts. It is good in translating specific domain area. It is not successful in literary texts.

MT uses various translating strategies, among which transfer strategy is mostly used in Arabic software programs. Transfer is a three stage strategy: where: 1) the source text grammar and lexicon is analyzed, 2) a transfer component is launched and 3) a synthesis component is produced.

Transfer systems permit taking into account syntactic sentence constituents in which lexical units appear.

4.3 Data preparation

This section prepares the scene for the corpora analysis.

4.3.1 Language Combination

In that the field of MT systems analysis is so broad, the scope of this study will focus on the single language pair: English - Arabic.

4.3.2 Text Types

Four different samples have been selected for machine translation and analysis: two medical texts; one is a medicine prescription, the second is an informative text about cold. A technical overview (Information Technology) and a news article (political). The first three samples were taken from the Internet and the political text was taken from the Gulf Today daily.
4.3.3 Users' Needs and Expectations

Users' needs and expectations depend largely on the sample domain. The MT users who translate the medicine prescriptions are either doctors, medicine salesmen or most probably patients who will use the medicine. In all cases, the need is to get accurate information about the medicine. Any mistake in this context could result in serious consequences for the users. The users' main aim of the second medical text is to assimilate information about the cold disease. The users of the Technical Overview need to have an overall idea about the main operational and functional ideas in the text. They need to have accurate information about "prototype tools from the web metrics test bed"... etc. The users expect accuracy and to have enough clear information. Users of the political text (a news article from the Gulf Today daily), to the contrary, need to get an overall idea of what is going on in the article The gist of the news may be enough for most political readers.

4.3.4 MT Systems

Two commercial systems were randomly selected from the market: 'Al-Wafi-version 4' developed by ATA company based in London, and 'Al-Kafi' developed by Cimos Company based in Paris. Both systems are developed to be used by, as it is mentioned on the software CD, translation centres, university students, newspapers and students studying in technical faculties. 'Al-Wafi' adds translators to the list.

4.4 Evaluation process

Since the evaluation procedure is based on MT users needs, this study does not venture into the technical and economic aspects of MT systems. Rather, it compares the quality of MT output using linguistic criteria in order to determine whether the systems do indeed satisfy users' needs.
Two types of criteria have been selected, one at the sentence level, the other at the text level.

1) Analysis will begin at the **sentence level** by checking: syntax, morphology and lexis.

2) The **overall text** will then be evaluated to check its readability and adequacy for the users' needs.

### 4.5 Analysis

In this section each TT translation output will be analyzed first as produced by Al-Wafi and then by Al-Kafi. The analysis will begin with two medical texts, followed by a technical text and finally a political text.

#### 4.5.1 Medicine Prescription/Medical Text

This is a medical prescription of anti-virus medicine used to treat Flu. It was published on the Internet by a manufacturing company in order to advertise this new product. The readership of such texts can be doctors, pharmacists or patients. In all cases, the translation is expected to be clear and accurate since it is essentially user oriented.

**BRAND NAME**: Symmetrel

**DRUG CLASS AND MECHANISM**: Amantadine is a synthetic (man-made) anti-viral drug that can inhibit the replication of viruses in cells. To prevent a viral infection, the drug should be present before exposure to the virus. Clearly, this is not practical for most viral infections. It was initially used to prevent influenza A during flu season, and, if given within 24 to 48 hours of the onset of flu symptoms, to decrease the severity of the flu. Later amantadine was found to cause improvement in the symptoms of Parkinson’s disease. Amantadine’s mechanism of action in Parkinson’s disease is not fully understood. Its effects may be related to its ability to augment (amplify) the effects of dopamine, a neurotransmitter in the brain, that is reduced in
Parkinson’s disease. Amantadine is less effective than levodopa in Parkinson’s disease but can offer additional benefit when taken with levodopa. Amantadine is less effective than levodopa in Parkinson’s disease but can offer additional benefit when taken with levodopa. Amantadine was approved by the FDA in 1966.

GENERIC AVAILABLE: yes
PRESCRIPTION: yes
PREPARATIONS: Amantadine is available as 100 mg soft gelatin capsules and as a syrup containing 50mg per each teaspoon.
STORAGE: Store at room temperature between 15 and 30°C (59 and 86°F).

PRESCRIBED FOR: Amantadine is used for the prevention or treatment of infections with influenza A virus, especially for individuals at high-risk such as immunosuppressed patients and nursing home residents. It should not be used as a substitute for vaccinating. Amantadine also is used for control of the symptoms of Parkinson’s disease.

DOSING: Amantadine is taken once or twice daily with or without food. If it causes an upset stomach, it can be taken with food.
For treatment of influenza, amantadine should be started within 24 to 48 hours after the onset of signs or symptoms. To prevent influenza, amantadine should be started as soon as possible after exposure to the influenza virus and continued for at least 10 days.
Persons with reduced kidney function and elderly persons may need lower doses (or less frequent doses).

Al-Wafi

The translated version

العلامة التجارية سيمبريل
Analysis

As previously mentioned, the analysis will be done at the three levels of: syntax, morphology and lexicon. Examples will be selected to demonstrate the problems when available.

Syntax

Reading through the translation output, the first thing which strikes the reader is that the translation is done using the word-for-word strategy. This type of strategy in translating creates a lot of linguistic problems on all levels, as Nida and others have said.
In literal translation, the translator usually sticks to the source text, but in accordance with the rules and confinesments of the target language. In word-for-word translation, the target text goes with the source text following all its rules and structures. Hence, the translation becomes odd and the message is usually lost. What makes things worse in the two translations of this particular medical text is that translation is done between two incongruent languages, i.e. English and Arabic.

1) As a result of the word-for-word strategy adopted by Al-Wafi to translate this text, word order and sentence structure appear corrupted if measured according to Arabic syntactic standards.

Examples,

- ......the drug should be present before exposure

  العقار يجب أن يكون حاليا قبل التعرض ...

- It was initially used to prevent ......

  هو كان أوليا يستعمل لمنع الآفات

2) Parsing: it is difficult to conduct parsing in such a confused sentence structure. It is impossible to consider cohesion here. However, and for the sake of checking the work of automatic parsers which the system developers claim they utilize, examples of parsing on individual cases will be examined.

In general, there is consistency in the application of Arabic syntactic rules between two consecutive structures such as VS or AdjN and others. However, there are cases of wrong parsing:

Example يجب أن يكون حاليا ... instead of يجب أن يكون حالي ...

In other similar case, the parsing was correctly done: كان أوليا ...
This indicates the lack of rules and consistency in parsing.

3) The system uses relative pronouns when they are not needed (they are used to follow the ST structure).

Example: - Amantadine is a synthetic drug ...drug that can inhibit ...

4) Consistency: consistency between masculine and feminine, adjectives and the nouns which they modify and between subject nouns and the main verb is preserved in general in terms of gender and number. However, there are still various cases of inconsistency.

Examples,

a) There is clear inconsistency in pronoun substitutions (in terms of gender):

- العقار يجب أن يؤخذ ... هذه ليست ...

- تأثيراته قد تتعلق بقدراتها على ...

b) There is a case of inconsistency between the dual verb and plural subject, and at the same time the modified noun and the modifying adjective forms are not following the same vocalization:

الأشخاص المسنون قد يحتاجون ... instead of ... الأشخاص المسنون قد يحتاجان ...

c) Inconsistency in using the definite article "ال" :

- تسبب الدوخة، التشويش، إغماء ...

It should be either

- تسبب الدوخة، التشويش والإغماء ...

or

- تسبب دوخة وتشويش وإغماء...
It is noticed that Al-Wafi system succeeded in some cases of syntax and parsing and failed in others. This is somehow strange since it is assumed that the parser and the syntactic analyzer in any machine translation system are built according to the rules of the TL so that the system respects such rules and the translation reads natural. Al-Wafi parser and syntactic analyzer must be more comprehensive in order to cover all Arabic rules. Consistency among verbs and subjects in term of number (singular, dual or plural) should be emphasized, for instance. Example sentences can be included in the built up of the analyzer for better application in the course of translation.

**Morphology**

Basically, the system has successfully identified and generated morphological variants of the nouns and verbs, especially in inflectional and derivational cases. However, there are cases where the same rules of Arabic morphology are violated.

1- Passive voice confirms the occurrence of inaccuracy in the translation output of this text.

Examples of passive verbs wrongly formed:

- "if given" is wrongly transferred into Arabic "معطى" instead of "أعطى"

- The verb "is used for" is transferred into "مستعمل" instead of "يستخدم".

- The imperative verb form store at room temperature ..., is again transferred into a passive "يمتزج" instead of "يغذى".

The passive voice in the above mentioned examples (case 3), should be extracted from the triradical (فعل + فعل) or (فعل + فعل + فعل) to have "يغذى" as اسم "معطى" and "يغذى" as فعل "يغذى".
2- There is also a problem in plural formation: In the sentence

- الأشخاص المسنين.. جرع أوطا...

The word جرع here is meant to be a broken plural of جرعة. But according to Arabic rules جرعة should take the natural feminine plural جرعتات.

However, all other plural forms are correctly generated according to the rules, even the broken plural like سكان, أمكان, أغراض and others.

It is clear that the morphological analyzer is enriched with most Arabic morphological rules. I believe, the rules of the passive voice needs more emphasis. However, such few mistakes and many others are natural in translated versions conducted by humans also.

**Lexicon**

There are many cases of mismatched lexical items in terms of semantics. These are lexical items which usually have various referential meanings, but whose usage differ according to context.

This is where the role of the human interaction plays an important part in selecting suitable meaning. It demonstrates the demand for understanding the pragmatic constituents where various technical items becomes clear here.

Examples:

1) ...drug that can inhibit the replication of viruses in cells. Replication is translated here as استنساخ. The intended meaning here is تكرار. Moreover, استنساخ has now acquired a fixed meaning cloning.

2) ....the drug should be present before exposure to the virus. Present here is literally translated as حالي. What it is intended here is يؤخذ..
3) ...to decrease the severity of the flu. Decrease is translated as نقص, while what is meant is.

In brief, the word-for-word translation of the previous 'text' is a good model of what Hutchins calls "unnatural literalness". The translation closely adheres to the source language structure and hence, it is generally odd. However, the user can get the gist of the meaning, if this is what he is looking for. H/She can get a general idea about the medicine. Nevertheless, it would be advisable in such texts to access other translations or to consult an expert in the field.

Al-Kafi

The translated version

The class and mechanism of the antiviral (صنع) نقص, which is the treatment for the flu. Decrease is translated as نقص, while what is meant is.

In brief, the word-for-word translation of the previous 'text' is a good model of what Hutchins calls "unnatural literalness". The translation closely adheres to the source language structure and hence, it is generally odd. However, the user can get the gist of the meaning, if this is what he is looking for. H/She can get a general idea about the medicine. Nevertheless, it would be advisable in such texts to access other translations or to consult an expert in the field.

FDA 1966

AVAILABLE

Name: Levodopa

Available by prescription via levodopa or levodopa as part of a combination.

Class: Levodopa and amantadine

Stability: Levodopa is stable at room temperature.

Available via levodopa or levodopa as part of a combination.

Class: Levodopa and amantadine

Stability: Levodopa is stable at room temperature.

Available via levodopa or levodopa as part of a combination.

Class: Levodopa and amantadine

Stability: Levodopa is stable at room temperature.

Available via levodopa or levodopa as part of a combination.
Analysis

The translation output of the Al-Kafi poses severe linguistic problems. Additionally, apart from linguistic problems, there are other problems which further hinder message and meaning. First, the system does not follow the Arabic right side text alignment. Second, the system is unable to read and 'understand' words, so in many cases the words are either written in English or transliterated using Arabic characters.

Syntax

Linguistically speaking, the whole sentence structure is confused. No system is followed in translating. It is not even a word-for-word translation. There is no clear word order and the structure of the sentences are either mixed up, or the sentences are incomplete.

Examples of the lack of structure in word order:

- يوضح، لم يكن هذا عملًا لأكثر عددًا فيروسية الفلوتزونا أثناء الفلوتزونا الفصل.
- كان مستخدمًا مبدئيًا أن يمكن إلى تحقيق سبب الهجوم أعراض الفلوتزونا.
However, for the sake of checking the efficiency of the syntactic and morphological analyzers, it is useful to examine the syntactic and morphological variants out of sentence; i.e., as individual cases.

1) Parsing: various cases of parsing are generally correctly conducted: example

- يجب على العقار أن يكون حاضرا.
- يعمل شخص ذاتا كلية...

2) There is consistency between the masculine and feminine modifiers and modified:

- عقار فيروس يقدر أن يمنع...
- تتمي مؤثراتها إلى قدرتها على...

Morphology

The system proved successful in the formation of inflectional and derivational words from their roots or stems.

Examples:

1- (فيروس، فيروسية، فيروسات، فيروس). Though the word is not originally Arabic, it has acquired the derivational rule of Arabic words.

2- From the tri-root (علاج)، the following words were derived علاج، معالجة، علاج

3- Broken and natural plurals are identified: علامات، جرعات، أمراض

Lexicon

The system poses serious lack of 'understanding' of the word meaning.
1- In some cases even the referential meaning is lost. Example: the replication of virus is transferred as جواب فيروس.

2- The collocation is strange (e.g. upset stomach is transferred معدة منزعجة).

3- A phrase like elderly persons may need lower doses is transferred قد يحتاج شيخ جرعات منظا.

These are few examples of so many strange and confused meanings and structures.

In brief Al-Kafi's translation of the medical text is very bad. It is unreadable, inadequate and unusable. The user can get almost no idea about the medicine in prescription, and probably would not continue reading after the first two sentences.

4.5.2 Second Medical Text

This is another medical text which is informative in nature. It is written for average readers and its aim is to expose some points of interest in regard to a topic of concern to most humans; cold.

The title of the article is How Colds are spread. It is published on the Common Cold Inc. site.

The site's aim, as announced, is to inform the public about colds, how they spread, their causes, symptoms and treatment.
How Colds Are Spread

Cold viruses grow mainly in the nose where they multiply in nasal cells and are present in large quantities in the nasal fluid of people with colds.

Highest concentration of cold virus in nasal secretions occurs during the first three days of infection. This is when infected persons are most contagious.

Cold viruses may at times be present in the droplets that are expelled in coughs and sneezes.

Nasal secretions containing cold viruses contaminate the hands of people with colds as a result of nose blowing, covering sneezes, and touching the nose. Also, cold viruses may contaminate objects and surfaces in the environment of a cold sufferer. Young children are the major reservoir of cold viruses and a particularly good source of virus containing nasal secretions.

Experiments have demonstrated that a cold virus readily transfers from the skin and hands of a cold sufferer to the hands and fingers of another person during periods of brief contact. Also, cold viruses readily transfer to the hands as a result of touching contaminated objects and surfaces.

Virus on the fingers is transferred into the nose and eye by finger-to-nose and finger-to-eye contact. Virus deposited in the eye promptly goes down the tear duct into the nose. Once in the nose, a cold virus is transported by mucociliary action to the adenoid area where it starts a cold. In some instances, cold virus, which is expelled into the air in coughs and sneezes, may land in the nose or eye and cause infection.

When the reader decides to translate this text using a machine translation software system, the goal here, according to Hutchins, is to assimilate information. In this case, the main concern is the message. The reader needs to gain the gist of information and he does not care much about the "naturalness" of translation and the aesthetics of the text.

However, examining the systems' successes and failures is necessary here to serve the objectives of the this part of the thesis.
The Al-Wafi has demonstrated relative success in translating this text. The passage is readable, it follows the rules of the Arabic language (except for some cases) and the information is to a good extent clear.

For the sake of system's evaluation, some examples of drawbacks will be selected for exposure.
Syntax

1) The misunderstanding of adjectives as verbs. Examples:
   a) infected persons is translated as أصاب الأشخاص instead of أصاب الأشخاص المصاصبين.
   b) contaminated objects is translated as الأجسام الملوثة instead of الأجسام الملوثة.

   Whereas the cold sufferer is translated as معاني بارد instead of الشخص المصاب ببرد.

2) The system followed the SL usage of particles where a verb or a noun should be used instead in the TT. Examples:
   a) ...fluid of people with cold is translated as السائل الأنفي من الناس بالبرد...instead of السائل الأنفي للأشخاص المصاصبين بالبرد.
   b) ...to the adenoid area where it starts a cold is translated as إلى المنطقة القرآنية حيث تبدأ برد...instead of إلى المنطقة القرآنية حيث تبدأ الإصابة بالبرد.

3) There is inconsistency among conjunctions. Examples:
   a) حيث يتضاعفون ويتواجدون instead of حيث يتضاعفون ويتواجدون.
   b) يغطون العصعص ويسمون الأذن instead of يرتفعون العصعص ويسمون الأذن.

4) The passive voice is used where the active voice is needed. Examples:
   a) ...where they multiply is translated as حيث يتضاعفون instead of حيث يتحولون.
   b) ...a cold virus readily transfers is translated as فيروس بارد يتحول بسهولة...instead of فيروس بارد يتحول بسهولة.

The opposite is used in the title. The verb in the English text is passive: How Colds are spread, whereas the verb in the Arabic text is transferred into كيف البرد انتشر.

Such syntactic problems indicate the need not only to expand the rules of the syntactic analyzer, but also to verify how these rules are used by employing the example-based strategy. For
exam ple, as for the consistency among conjunctions, the rule is definitely included because it was used in other cases, but it needs enhancement through examples to demonstrate application.

**Morphology**

The system succeeded to follow the morphological rules of Arabic in the formations of: natural and broken plurals (such as الفيروسات, الأفرزات; الأيدي; الأصابع), the superlative form (such as الأعلى), the formation of verbs according to their patterns and their position in the sentences (تتمو، يضايفون، بيئت، أصاب) (keeping in mind how the systems 'understands' the verb) and the formation of passive forms (such as مترود، مصطلح، موجود).

**Lexicon**

In some cases the system failed to catch the meaning of some words and expressions whereas in other cases the selection among the synonyms is not accurate. Examples:

1) Cold viruses is translated as الفيروسات البرد instead of الفيروسات المعرفة.

2) *...droplets that are expelled in the coughs* is translated as القطرات السطرودة instead of التي تطرح مع...

3) The word eye is sometimes correctly translated as عين and other time it is translated as يرى.

4) *...children are the major reservoir of cold viruses* is translated as الأطفال الصغار الخزان الرئيسي الخزان. The word الخزان is not a good selection in this context. A better selection, I believe is المخزن الأساسي لجمع...

5) The expressions: finger-to-nose and finger-to-eye contact is wrongly transferred as بالاصبع للاشتمام وليس إلى الإتصال العيني. And once in the nose is also wrongly transferred as مرة في الأنف.

6) The system failed to 'recognize' the meaning of the medical term mucocilliary and kept it in English.
On of the problems of meaning, I assume, is that the database include one derivational meaning, as it is the case with the word جر. If other derivations were included and some examples of usage were used, this problem would be easily solved.

In general, the text in general is readable, adequate to the needs of the reader whose aim is to assimilate information, and it reads natural to a large extent. Although the above examples indicate the existence of drawbacks in translation, yet the message is relatively clear. With quick post-editing the text becomes eligible to publication.

Al-Kafî

The translated version

كيف انتشر كولدس اربي

فيروس بارد تم في الغالب في الأنفل أن يتضايقون في خلايا أنفية وحاضرون داخل كميات كبيرة في مائع الناس الألف مع برد.

يقدر عدد أعلى فيروس بارد في مفرزات الأنف أثناء أيام الثالثة الأولى عمود. هذا عندما أشخاص ملوثون هم أكثر معدن. ربما قد يكون فيروس بارد أحيانا حاضرا في القطب التالى أنفية الذي في سعالان ومعطيات مفرزات أنفية التالى

مفرزات أنفية تشمل بارد فيروس بلوش أدي أدنى مع برد شعاع لأنف يشبه، تعطى فقط، والتي تتم الأنف. أيضا ربما قد يلعب فيروس بارد أشياء ويتضمن في بيئة معتمي بارد. أطالف بونغ هم خزان الفيروس الباردة الرجية ومصدر جيد خاصة الفيروس الذي تصل مفرزات أنفية.

يجب جد كشف عن ذلك فيروس بارد بسرعة تحويلات من الجلد وأدي معتمي بارد إلى الأيدي وأصابع الآخر الشخص أثناء فترات الاتصال الموجز. أيضا فيروس بارد بسرعة أرواح تحول إلى الأيدي كما للس أشياء وسلطق.
Analysis

The translation of the text demonstrates similar drawbacks as it was the case in the previous translated versions, with slight improvement in certain places where the structure of Arabic language is followed. However, it is still difficult to examine the TT structure since the word-for-word policy was adopted in the translation of text without even respecting the TL linguistic rules. Syntax is one example.

Syntax

It is impossible to follow the syntactic rules of Arabic here since the whole structure of the text follows the English structure. Parsing of course is impossible since there is not clear structure. There is no consistency for example, in parsing among the modifiers and the modified (such as the person's name). The system failed even to build simple structures such as (it is used as the person's name) and (the person's name is used as the person's name).

There is only one sentence which the system could build according to the Arabic structure (VSO): ...

فیروس على الأصابع ينتقل في الأنف وعين بواسطة أصبع-الأنف وأصبع-عين اتصال أودع فيروس العين داخل بحزم يحبط الدمعة قناة في الأنف، ذات مرة في الأنف، ينتقل فيروس بارد إلى المساحة الخفيفة في النطاق. في بعض الحالات، فيروس بارد بواسطة عمل يدوي الذي في الهواء في ساعتين وعشر، ربما قد تحت في الأنف أو يندفع إلى وسبب عدوى.

Analysis

The translation of the text demonstrates similar drawbacks as it was the case in the previous translated versions, with slight improvement in certain places where the structure of Arabic language is followed. However, it is still difficult to examine the TT structure since the word-for-word policy was adopted in the translation of text without even respecting the TL linguistic rules. Syntax is one example.

Syntax

It is impossible to follow the syntactic rules of Arabic here since the whole structure of the text follows the English structure. Parsing of course is impossible since there is not clear structure. There is no consistency for example, in parsing among the modifiers and the modified (such as the person's name). The system failed even to build simple structures such as (it is used as the person's name) and (the person's name is used as the person's name).

There is only one sentence which the system could build according to the Arabic structure (VSO): ...

فیروس على الأصابع ينتقل في الأنف وعين بواسطة أصبع-الأنف وأصبع-عين اتصال أودع فيروس العين داخل بحزم يحبط الدمعة قناة في الأنف، ذات مرة في الأنف، ينتقل فيروس بارد إلى المساحة الخفيفة في النطاق. في بعض الحالات، فيروس بارد بواسطة عمل يدوي الذي في الهواء في ساعتين وعشر، ربما قد تحت في الأنف أو يندفع إلى وسبب عدوى.
There are some cases where the system could achieve consistency for example (taking into consideration the consistency of gender regardless whether the derivations are correct or not).

**Morphology**

The system could achieve some success in the formation of some morphological structures like verb and noun formations, such as یتضااععون، يقع، يث旅行社، كميات، تحويلات، معاني، أيدي. However, the formation of some verb and noun forms seems very odd.

Examples:
1) اسم вместо اسم.
2) سعال вместо سعال.
3) أفرزات вместо أفرزات and
4) أنفيات вместо أنفيات.

**Lexicon**

The worst part of the translation of this text is the translation of meaning. Although the text is medical, yet it has very few medical terms. The language used is not a jargon, it is a simple language since the text addresses the public. Yet, the system failed to 'recognize' the meaning of a large number of words, and if the meaning is there, the selection of the synonyms is not accurate.

Examples:
1) ...nasal fluid of people is translated as السائل الأنفي لدى ... instead of ...
2) ...infected persons is translated as أشخاص مصابون instead of ...
3) cold virus is translated as فيروس بارد.

4) ... a cold virus readily transfers from the skin is translated as فيروس بارد بسرعة تحويلات بتنقل فيروسات البرد بسرعة instead of تركز عال حامد.

5) Highest concentration is translated as حشد instead of تحط.

6) In the last sentence, the word land in cold virus ... may land in nose... is translated as تحط. The translation is correct, but the Arabic synonym is too strong in this context. تستقر في الألف is a better translation, I believe. 

The system astonishingly could not recognize simple words like young in young children. The word is transliterated as بنتنج. Another example is the transliteration of the colds are in the title as بولدس آري.

5) The medical term mucocilliary is kept as it is in English. This term is a compound of mucoco-cilliary. This term is not recognized by both Al-Wafi and Al-Kafi.

The failure to give the meaning of this term indicate the need include medical databases which cover medical prefixes and suffixes since a great number of medical terms are formed through compounding.

In general Al-Kafi has failed in most cases to respect the linguistic rules of Arabic language. What is astonishing in fact is its clear failure in the lexicon part for various reasons: first, the system is supposed to include a variety of dictionaries, among them the medical lexicon. Second, the system supposedly includes databases that contain references for thousands of words. The words of this texts are simple and common and hence, they should be part of any database.
The text needs a great effort from a human translator or an editor to make it acceptable for dissemination. For information assimilation, the reader may succeed to get some ideas if s/he works harder to get the meaning out of the confused structure.

4.5.3 Technical Overview/Information Technology

This text is a technical overview. The title of the article is the *Usability Evaluation of The Website*. It was published on the Internet by the National Institute of Standards and Technology.

The readership here is not necessarily specialized experts in the field. It may include university students and individuals interested in this topic. The users are looking mainly for information.

Language is not a pivotal factor. However, the text must be readable in order to adequately meet the users' need; the assimilation of information.

Technical Overview

Good usability is critical to the success of a website. Usability evaluation has traditionally been a slow, labor-intensive process which makes it difficult to apply to websites. The dynamic nature of the Web poses problems for usability evaluation. Development times are rapid and changes to websites occur frequently, often without a chance to re-evaluate the entire site.

Advances in web-based user interfaces change user expectations. Finally, the potential audience for a website may be geographically dispersed and encompass a wide range of demographic groups.

The challenge then is to determine how best to provide automated support to the usability engineer. Automated techniques cannot entirely supplant manual testing; the intuition of a good usability engineer is still vital. However, automated techniques can enhance traditional approaches and provide additional information to the developer as well.

The objective of the NIST Web Metrics Testbed is to explore the feasibility of a range of tools and techniques that support rapid, remote, and automated testing and evaluation of website usability. The prototypes are used to support the usability engineering research of the Visualization and Usability Group (VUG). As part of the Information Access Division of the Information Technology Laboratory at the National Institute of Standards and Technology, VUG encourages industry to use and/or commercialize its ideas.

The NIST Web Metrics testbed was undertaken in 1997 to explore solutions to the problems described above. We released version 1 (consisting of simpler versions of WebSAT, WebCAT,
and WebVIP) in June of 1998. Since that time, we have enhanced the original software and added FLUD and VisVIP.

We are now (October 2002) releasing a new version of Web Metrics with even more improvements. Please take a look at our What's New page for the details.

Prototype Tools from the Web Metrics Testbed – DOWNLOAD

The Web Static Analyzer Tool (WebSAT) checks the HTML of web pages against usability guidelines, either its own, or a set of IEEE Std 2001-1999 guidelines. It can check individual pages or an entire website.

The Web Category Analysis Tool (WebCAT) lets the usability engineer quickly construct and conduct a simple category analysis across the web. It is a variation upon traditional card sorting techniques. The usability engineer establishes a set of categories and a number of items which are to be assigned by test subjects to those categories. The engineer can then compare the actual assignment with the intended usage to make sure that the categories match users’ intuitions.

Al-Wafi

The translated version

النظرية العامة للتكنولوجيا

إن التحدي ثم أن يحدد كيف أفضل لتزويد الدعم الفني إلى مهندس usability وكالة اختيار بدوية كلما، لدى مهندس usability بسيط ما زال حيويًا. على أي حال، اعتقد الاستاذ تستطيع تحضير نظرة تقنية وترقى معلومات إضافية إلى المطور أيضًا. ميترس تستفيد أن يستكشف عملية مجموعة من الأدوات وإمكانيات التغيير الاستاذ usability وبناء علي المعركة المقبولة على الشبكة العالمية VUG من فيسوايزيشن و مجموعة واسعية VUG. كجزء من قسم وصول معلومات مختبر تقنية المعلومات في المعهد الوطني للمعايير والتقنية، يشجع VUG صناعة لاستخدام و/أو يتاجر بأدواته.

The Analysis

The translation output of the technical overview will be examined to see how far the translation serves the above-mentioned users' needs.

Syntax

The translation strategy used again in translating this text is the word-for-word strategy. Many problems will raise consequently.

1) It is very clear right from the beginning of paragraph one that sentences are scrambled. They have no clear structure, and many of them are incomplete. Moreover, it is difficult to trace a clear word order even on the basis of word-for-word strategy. Examples:

1- أجل، كانت تعديلات السيرة تختلف المواقع الويب تحديداً.
2- قد يتعلم دواعي مساعدة مستعمل التوقفات على الإنترنت.
2- حسن مهندس، جيد ما زال حيوي.

2) It is difficult to examine parsing in such confused 'structure'. Since there is no clear sentence structure or sentence segmentation, parsing becomes an impossible task.

However, there are few cases where the system can 'recognize' Arabic syntactic rules (out of sentence environment)
a) When there are cases of (VS) structure, or full (VSO) structure for example, the system in some cases applies the rule of Arabic parsing.

Examples:

b) The modifying adjectives carry the same noun diacritization.

Examples:

c) Consistency is achieved between the modifiers and the modified variant items in terms of masculine or feminine.

Examples

d) The word order of the title is confused. There is no need for the definite article (ال). The title should be translated as نظرة تقنية عامة. However, the meaning is clear.

Morphology

In general, the system follows Arabic morphological rules.

Examples:
1) Words like تقييم and تقدم are formed from the root (فعل) as (قيم) and (قدم) consecutively.

2) The system can generate derivations like مجموعات, جمع, عملية, and Embed, عمل. Others like نماذج, نموذج, أساليب, أساليب, and others.

3) The broken plural is recognized in the system and others.

4) Other inflectional forms are generated, such as استكشاف, يكتشف.

**Lexicon**

1) The key word of the text is usability. It was in some cases translated as حلول, in other cases it was transliterated as واسطويتي, and in some other places it was kept as it is in English.

2) There are words which the system 'failed' to 'recognize', so it 'decided' to use transliteration strategy, though they are key words in the text.

Examples,

metrics is transferred and testbed is transferred كيستيبيد (one word instead of a compound term).

3) Some words are sometimes translated and in other times transliterated.

Example: website. It was first translated as شبكة العالمية, and in other cases, it was kept as الويب.

4) There are key words which the system failed to 'understand', and they were given different meaning. Example: critical in the first sentence. The word critical here means very important.

The system transferred the word as نادر. The problem here is mainly of pragmatics nature. The system needs to 'enjoy' world knowledge or even 'common sense' to 'realize' the differences in meanings.
5) When it comes to the meaning of the two prototype tools, the translation in general is not very clear, but with little post-editing, it becomes understandable.

6) The technical terms of the prototype section are transliterated. This strategy can be helpful for the Arab users who know the terms only in English. But on the other hand, the transliteration of terms means that Arabicization is hindered.

Failure to translate the technical terms demonstrates the need for regular expansion and updating of technical dictionaries, databases and encyclopedia.

The previous cases clearly represent what Vinay and Darbelnet categorize as "unacceptable message" in translation due to literal translation (item h).

Al-Kafi

The translated version

تدشنيكال أولفر فاير

قد كان تقييم صلاحية تقنيات بضاء، شلل - مركز عملية. صلاح جيد اقتناطي إلى النجاح لـ website. تتفاوت طبيعة الويب ديناميكية المشاكل تقييم. الهدف يجبره صعبا أن يطبق على طالما، غالباً الصلاح. أوقات تطور سريع و تغير تغييرات على إعداد - قيم الموقع website بدون صعوبة إلى الكامل، برئي في راهن مستعملة مرتكزة على الويب توقعات مستخدم تغيير ربما قد يفرق جغرافيا و أخيرا، جمهور المستعنين الكآمن ل ديموغرافية عريضة.

سيحدد الرحلة ثم كيف أفضل أن يزود شغل اتوماتيكي سند إلى الصلاح مهندس. شغل اتوماتيكي تقنيات يقدر بمستند كلية اختبار، بدأ ببداية مهندس صلاح جيد حبيبي، وم، مع ذلك، شغل اتوماتيكي تقنيات يقدر يحسن طرقا تقليدية و يوفر معلومات إضافيات إلى المطور كذلك إن أي إست تي سيكتشف هذه website ميترس تيصيف ويبي معقولية سلسة أدوات و تتحاول تقنيات إلى النماذج الأولية. هذا الحد website بسرعة، بعيد، و شغل اتوماتيكي اختبارا و تقييم صلاح (مستخدمة أن تتحمل الصلاح التي تهندس بحث فايز والبيتانيون و سالينيتي جروب (في بحوي جزيء التفريعات كبس وحدة قسمة إنفور ماتيون تيتشنولوجي لاباراتوري في نانويال أسيستيتوي لستانداردس و تيتشنولوجي، شجع في بحوي صناعة إلى استعمال و/أب يتأخر أفكار ما.
Analysis

Reading through the Arabic translation of the Technical Overview translated by al-Kafi, the reader feels lost. There is no sentence structure. Sentences are formed by words put together in an unsystematic manner. They are not linked, hence cohesion is lost. The translation is not even a word for word translation.

Syntax

To examine the syntactic structure of the Al-Kafi translation of the Technical Overview is an impossible task. The translation output is anything but a text to be read and understood. However, some sort of syntactic analysis will be conducted to see if the system can adhere at least to certain cases of syntactic rules.

1) Word order. There is no word order organization in the translation output.

Examples,
2) Sentences are incomplete. Examples,

- أوقات تطور سرعات وتوقع تغيرات إلى إعادة قيم الموقع;
- تركل بديهة مهندس صلاح جيد.

3) Punctuation is used randomly. Examples,

- قد كان تقييم صلاح تقليديا ببطء ، شغل ، مركز عملية.
- ووقع تغيرات إلى إعادة - قيم الموقع كامل.

4) It is impossible to apply any parsing in such sentences. However, the first sentence قد كان تقييم صلاح تقليديا is correctly structured according to the rules of the weak verb كان.

Morphology

1) The key word of the whole text is usability. It was translated as صلاح . It seems the derivation is based on the pattern ( نورل ), which is very odd here. The acceptable derivation here might be صلاحية.

Other morphological formation rules are generally applied well. Examples,

- منشأ - نشأ
- إصدار - أصدر
- استكشف - استكشف - كشف
- أوقات - وقت
- لشكل - شكل
- مستمع - مستمع
Lexicon

1) The system failed to 'recognize' the referential meaning of some key words such as: critical
t(اتقادي)
2) Transliteration strategy was used in many cases:
   a) the title Technical Overview was transliterated as تيتشنيكال أوف سو.
   b) Proper names are like فايزو والياتيون و وساليتي جروب for Visualization and Usability Group.
   c) Technical terms such as testbed (تيسبيدي), metrics as (سيريس).

However, website, which is widely used as شبكة عالمية is kept in English.

In short, this text is unreadable and inadequate. It is much easier to re-translate the whole text
than to try to post-edit it.

4.5.4 News Article/Political

This news article is published in the Gulf Today daily on the 22nd of April 24, 2004. The title of
the article is Arabs describe Riyadh attacks as barbarous. The article handles Arab stands
towards recent suicide attacks in Saudi Arabia. The assumed reader here needs to have an
overall idea about the Arab stands towards such attacks in a brotherly country. The possible
readership here is an Arab who does not know English and who is either in a foreign country or
on board the plane where only English news papers are available, or otherwise, he wants to
know how English news papers tackle such topics. In order to get the information, s/he uses
machine translation for quick translation. His/her main aim is to get the information s/he is
looking for no matter the standard of translation.
Arabs describe Riyadh attacks as barbarous

RIYADH: Arab states on Thursday condemned as a "criminal" act the suicide car bomb in the Saudi capital which killed at least four people and wounded 145, and said the attack violated Islamic principles.

"We condemn this criminal and terrorist act against a building of the security agencies in Riyadh and we express our condolences to the families of the victims and hopes for a speedy recovery of the injured," Syrian President Bashar Al Assad said in a message to King Fahd of Saudi Arabia, according to the Sana news agency.

The highest Islamic authority in Syria, Sheikh Ahmad Kafiaro, also denounced the bombing as "barbarous and inhuman".

"Those who commit barbarous and inhuman acts are very far from the spirit of Islam. In this way they serve the enemies of the (Arab) nation," the mufti said in a public message.

Kuwaiti Prime Minister Sheikh Sabah al-Ahmed Al Sabah also denounced the "criminal explosion" during a telephone call late Wednesday with Saudi Crown Prince Abdullah, the official Saudi news agency SPA reported.

In Doha, a Qatari foreign ministry spokesman said such "criminal acts go against the precepts of Islam and human and moral values".

Saudi US ambassador Prince Bandar Bin Sultan said in Washington that the attack was aimed at "the Saudi people and the royal family and officials of the government...are all Saudi citizens."

Bandar, who spoke on Wednesday after meeting US national security adviser Condieezza Rice, pledged his nation would "fight them (terrorists) hard" there will be no compromise.
Al-Wafi

The translated version

يصف العرب هجمات الرياض كبرى

الرياض: أفادت الدول العربية يوم الخميس كعمل "إجرامي"، عملية تجبر السيارة الإحتلالية في العاصمة السعودية التي عبّلت على الأقل أربعة أشخاص، وجرح 145، وقالت الهجوم إنها إهانة إسلامية.

الذي هذا الهجوم والعمل الإرهابي ضد بداية وكالات الأمن في الرياض ونحن نيدي تذكيرا إلى عوائل الضحايا والآمال في العالم العربي.

الرئيس السورمي بشار الأسد قال في رسالة إلى الملك فهد عاهل العرب السعودية، طرقا لوكالة نيويورك.

الناتجة الإسلامية الأعلى في سوريا، النتيجة أحد كافئنا، شجب التصعيد أيضا ك"برء ولا إنساني".

"والندوة الذين يرتكبون الأفعال الإرهابية ولا إنسانية بعيدا جداً من روح الإسلام. بهذه العطتها يخدمون أعداء (عربي) آمنة.

"المدي نقل في رسالة عامة.

شجب رئيس الوزراء الكويتي الشيخ صباح الأحمد الصباح أيضا "ال отлично Ejragmi" أثناء مكالمة هاتفية في وقت متأخر من يوم الأربعاء مع ولي العهد السعودي الأمير عبد الله حمام وكالة الأنباء الرسمي السعودي الرسمي دقائق.

في الدوحة، نقل رئيس وزارة الخارجية قطري قال: "مثل هذه الأعمال الإرهابية تضمن تضامن الإسلام والسعودية والأخلاقيات.

الناتجة الأمريكية السعودية الأمير بدر بن سلطان سيف في والقين الانتهاك الذي الهجوم إعتدنس "العثور السعودية والعائلة المالكة ومسلت Colonizer رايس، تعهد بأنه "تخريج" (إرهابيين) بشدة "أن لن يكون هناك مساهمة.

Analysis

Reading through the translation done by Al-Wafi, it is noticed that the standard of translation is very close to the standard of good human translation. The text in general reads Arabic. It follows the syntactic and morphological rules of Arabic language to a large extent. The message of the text is clearly expressed and the selection of Arabic synonyms is successful. However, there are
few cases where the choice of words can be more accurate, some structures are more English
than Arabic and few other weak points.

The analysis will concentrate on the drawbacks only since the majority of the text is good.

Syntax

The sentence structure of the TT follows the Arabic rules except for few cases.

Examples: يخدمون أعداء الأمية العربية. The correct structure is يدؤون أعداء الأمية العربية...

2) There is one case of inconsistency between conjunction particles: ونأمل الشمال العامل للجنحي والجمال لشنه عامل لجنحي. The last part should be بناه الشمال العامل للجنحي.

3) The addition of certain words to some sentences can improve the translation.

Example:

a) The title reads باثها بربرية instead of يصف العرب هجمات الرياض كبريرية.

b) The text says "يمعبرا إيابا عملا .. واصفا .." شجب القصص أيضا ك "بربري ولا الناسي". If a phrase like "يمعبرا إيابا عملا .. واصفا .." the translation will be perfect. This structure is repeated many times.

b) The last sentence reads "تتعهد بأن تحارب أعداء الأرهاشين بشدة". The right structure, I believe, is "تتعهد بأن تحارب أعداء الأرهاشين بشدة ولا تكون هناك مسومة.

Morphology

The system has fully succeeded to comply with the morphological rules of Arabic language. No failure is noticed.
Lexicon

1) There are few weak choices of verbs which should be replaced with verbs which collocate better in the context.

Examples:

a) ... وقالت الهجوم انتهك...، a better choice is. 

b) ... هذه الأعمال الإجرامية تصرف ضد...，a better choice is ...

2) The system made a very critical mistake when it confused the abbreviations of the Saudi News Agency SPA as حمام وكالة الأخبار الرسمي.

3) The system strangely failed to 'recognize' the last name of Saudi Prince Bander bin Sultan. It transferred Sultan as سولتان، although the name is very popular in Arabic. The verb said is also transliterated into Arabic as ميد.

The last two points indicate the shortage of database for both acronyms and proper names. All in all, the translated version of the political article reads natural. And the message is clear. The text can go for publication with minor improvements. It worth noting that no human translator can achieve this result in less than two seconds as the system did.
Analysis

Again Al-Kafi has failed as appears from the translation to achieve a good readability and hence, it failed to give any message. The word-for-word translation strategy and the adoption of the English structure made the whole text a failure.
Syntax

There is no need to again demonstrate the system's failure in regard to sentence structure since the same policy was used in the translations of the previous texts where the odd structures resulting from such a policy are already exposed.

Of course parsing is another syntactic feature which is impossible to examine since the sentence structure of the text is odd to Arabic system.

Morphology

Morphology is usually the successful part of both systems, and it is the only successful part of Al-Kafi. Derivational and inflectional formation of words were done according to the morphological rules of Arabic. However, there are very few mistakes.

Examples:

1) The formation of the broken plural of ناش is oddly formed as ناشة.
2) The noun سرعة is again oddly formed as سرعة in ... وتشتت سرعة استرداد.

Lexicon

Opposite to morphology, lexicon is the part of the system which demonstrates real problems in spite of the fact that each system is essentially built on data bases which enrich the system and provide the required vocabulary and terminology for translation. Al-Kafi usually fails to 'recognize' meanings and hence it goes either for transliteration or it keeps the words as they are in English.
In this text, the system mainly failed to 'recognize' the names and the titles of the Arab leaders. This is a critical mistake because these names and titles are common and consist part of everyday news in newspapers, TVs, and radios.

Examples:
1) President Bashar Al-Assad is transliterated as بشار الأسد.
2) Prime Minister Sheik Sabah Al-Ahmed is transferred as الشيخ صباح الاحمد.

In other instance, the telephone call is transferred as تليفون داه.

However, the system could 'recognize' the name of the Saudi News Agency SPA which Al-Wafi failed to 'recognize'.

It seems that Al-Kafi also lacks databases for proper names and titles. It is important to include such information to avoid easy and clear mistakes.

It is concluded that the Al-Kafi system developers need to improve and expand the scope of the syntactic analyzer especially in terms of sentence structure which constitute the major problem in the translations conducted by the system. If this problem is solved, then translation would be more natural, texts would be more readable and accordingly the message of texts becomes clearer. In regard to morphology, the system, it seems has employed good morphological analyzers. Although morphological analysis is not an easy process, it proved successful contrary to a supposedly more easier task; lexicon. Lexicon is expected to prove the most successful part
of any translation system since the systems essentially depend on a large number of various
dictionaries, databases and translation memories.

In short Al-Kafi needs to re-evaluate its whole system of translation. The standard of translation,
as it appeared from the translation of three different types of texts is very poor, unnatural,
unreadable and inadequate to the users' needs.

4.6 Conclusion

Given the analysis conducted in the previous section, the following conclusions can be made:

1) Arabic machine translation software developers claimed that they adopt transfer as a
translating strategy. This means that translation is done on three levels: the source text is
analyzed and transferred into an intermediate language called a meta-language with the help of a
TL lexicon and then restructured before transforming the sentences according to the syntax of
TL (Hutchins, 1986). However, the previous corpora analysis demonstrates that Al-Wafi used a
word for word and the literal translation strategies in their translations.

Al-Kafi failed even to follow the simple straightforward translating strategy; word-for-word.
Al-Kafi's output is merely a combination of words put together randomly without a strategy or a
structure.

2) Subsequent to the kind of strategy adopted, word order in most cases did not comply with the
Arabic code structure (basically VSO). Relative improvement appeared in the translation of the
second medical text and the political text by Al-Wafi. In Al-Kafi, word order is mostly confused.

3) With regard to syntax, one cannot talk about cohesion even on the sentence level. However, the two systems might achieve a good ability to deal with syntactic phenomena like consistency and dependency between the variant forms of modifiers and modified norms (keeping in mind that this possible only for the sake of analysis out of text, i.e. as individual cases).

4) Both systems, Al-Kafi and Al-Wafi, demonstrated very good ability to analyze and generate forms of Arabic words according to the rules and structures of Arabic morphological rules (such as derivational and inflectional rules).

5) With regard to meaning, both systems succeeded to give the referential meaning of a number of ST words. However, the two systems demonstrated cases of tremendous failure in 'realizing' certain words which: first, hold various meanings and second, whose meaning depends largely on the understanding of the context, or what the linguists call the world knowledge or pragmatics.

In short, the translation of the medicine prescription done by Al-Wafi is readable to a certain extent. The reader can get the gist out of it, though this is not enough in such texts because if they are not clear and accurate enough, the translation may cost the patient his/her life. The second medical text demonstrated improvement both on readability and meaning.
Regarding the technical overview, the user would most probably not understand much of the first part of the overview because the language of the ST was not straightforward because the system could not realize the meaning very well. However, the user would be able to get some idea about the prototype tools from the website metrics part because the translation here was clearer and more accurate. This is maybe due to the fact that this section is more technical and the language is clearer.

The political news article was the best translated of all. The language is readable, the meaning is evident and it needs only quick post-editing to make the translation perfect. It appears easy to achieve meaningful 'text' in this case because the language used is simple and common.

As for Al-Kafi, the translation is poor and unreadable and hence the meaning is almost lost. Post-editing does not work here.

By applying the results of the analysis to what has been provided in the introduction, in terms of linguist viewpoints regarding translation, one can notice that:

1) As is mentioned in (item b, pp.95 - 96), Catford is correct in his theory of contextual meaning and the fact that this approach is still very recent in MT. It is clear from this analysis that the lack of 'understanding' of the context, or world knowledge is applicable to the MT systems of all languages. However, it appears even more severe here since the translation strategy followed in Arabic MT translations is largely word-for-word (especially by Al-Kafi).
2) The two translation strategies used by both Al-Wafi and Al-Kafi, when a strategy is available, are clear representations of Nida's examples of translating improvement from word-for-word strategy to literal translation (item d, p. 97). Nida's example follows:

- Greek ST:
  1 2 3 4 5 6 7 8
egeneto anthropos, apestalmenos para theou, onoma auto iōannēs
- Literal transfer (stage 1):
  1 2 3 4 5 6 7 8
became/happened man, sent from God, name to-him John
- Minimal transfer (stage 2):
  1 2 3 4 5 6 7 8
There came/was a man, sent from God, whose name was John
- Literary transfer (stage 3, example taken from the American Standard Version, 1901):
  1 2 3 4 5 6 7 8
There came a man, sent from God, whose name was John
or (example taken from Phillips New Testament in Modern English, 1958):
  2 6 7 8 3 4
A man, named *John was sent by God

Figure (10): Nida's model of translation improvement.

3) According to Nida's 'four basic requirements' for the success of translation (item f), all four metrics are lost in large in the translation output of Al-Kafi. In Al-Wafi, translation of some cases makes sense, the second medical text and the political text have instances of natural and easy form of expression, hence they produce almost similar response since the message is clear.
4) Vinay's and Darbelnet's indicators of the unacceptable message in translation (item h, p. 98) apply to the translation output of Al-Kafi. As for Al-Wafi, this applicable where the strict word-for-word strategy is followed.

4.7 Recommendations

1) It is recommended that in order to get acceptable results in Arabic machine translation, Arabic software systems must abandon the strict word-for-word strategy.

2) There is a need to improve and expand the syntactic analyzers and the parsing devices used in the MT systems to include all syntactic rules of the Arabic language.

3) In order to improve the applicability of such rules, examples from the Arabic literature, encyclopedia, newspapers and magazines and other sources should be supplied.

4) Dictionaries and databases should be expanded and upgraded on regular basis.

5) The systems should be supported with databases about acronyms, proper names, titles and other important information to abandon easily avoidable mistakes.

6) In order to improve systems' ability to 'recognize' the pragmatic meanings, AI strategies should be employed. In addition, rich databases and encyclopedia and the adoption of example-based strategy will help in this very complicated side of translation even to human translator.

7) MT system should adopt interactive translation strategy where human aid is supplied when necessary, either as pre-editing or post-editing during the process of translation.
Most MT systems currently developed are capable of translating scientific and technical documents. The translation of literary texts, as compared to technical texts through MT involves more complexity as regard to syntax and semantics. The literary language requires expressions for emotions and sentiments with much rhetoric and metaphors. Such translation demands human involvement so as to interpret the various literary intricacies of a literary language in order to produce meaningful translation.

For the one-to-one interchange of information, there will probably always be a role for the human translator, e.g. for the translation of business correspondence (particularly if the content is sensitive or legally binding). But for the translation of personal letters, MT systems are likely to be increasingly used. Likewise, for electronic mail and for the extraction of information from Web pages and computer-based information services, MT is the only feasible solution.

Today, the world has witnessed a changing context for machine translation. MT technology development has taken on broader significance in an age of rapid international communication and intense market competition. Competition in the global market has intensified the need for companies to sell their products to overseas customers who speak foreign languages. Some large companies have targeted translation technologies as a component of their competitive strategy. Another related explanation for changes in perspectives on machine translation is the information explosion. On a more practical level, there are also political factors in the search for good quality MT. In Europe, multilingualism is a fact of life, which makes translation necessary for communication. However, translation is time consuming and continues to be expensive, so MT could be a financial blessing.
Approaches in MT are very diversified. Some researchers see MT as a means of demonstrating their theories, with their measure of success based on whether or not the system is an accurate model of human mind or simply a 'pure' theory. Other researchers concentrate only on applying formulas lacking theoretical grounding. In fact, research in MT is still, above all, experimental but guided by solid theoretical foundations. Its sole performance criterion is to obtain results for a well-defined need. There is no global solution, however, for every translation need there is an adapted MT solution that considers the expected results and constraints on resources, cost and time.

Machine translation technologies pose a range of theoretical, software, hardware and even sociological problems that require the integration of technologies and improved interaction among developers and users. For these reasons, machine translation today is more than a linguistic problem. It is a communicative and informational challenge that demands a diverse range of expertise and resources.

The level of complexities of a MT system depends on the relative relationship between syntactic levels and other linguistic aspects of the source and target languages. In a direct translation strategy, a text is analyzed and is directly transferred into the TL through a series of stages of operations. The output of this system depends on a codified dictionary and the pre-specified sentence patterns and also on morphological analysis. In the case of the transfer method, the SL text is analyzed and transferred into an intermediate language called a meta-language with the help of a TL lexicon and then restructured before transforming the sentences according to the syntax of TL. In the case of Interlingua strategy, an intermediate or universal language is used...
for translation. Adopted for this method are Artificial Intelligence tools involving a high level structure and appropriate inference mechanism to resolve syntactic and semantic ambiguities and pragmatics.

The translation of a natural language is not just matching of words but is rather a conceptual transfer as opposed to a syntactical transfer. In order to design an efficient and usable MT system, it is imperative to analyze, interpret and understand the complex syntactic and semantic aspects of a NL. The major problems encountered during the MT process regards semantics rather than syntactics. It arises mostly due to the inadequate details of semantic representation and inefficient techniques adopted to represent the ambiguous situations and contextual variations. The most complex NL problems as related to MT are syntactic ambiguity, lexical and semantic ambiguities and idiomatic expressions, pragmatics or language in context, ellipsis, substitution and anaphoric references.

Fortunately, resolving such ambiguities is possible if we rely upon the interactive involvement of the user in what is known today as interactive systems. In these systems the user makes final decisions and resolves persisting ambiguities since no program is able to integrate sufficient world knowledge and common sense so as to automatically resolve all of the ambiguities in any source text for many years to come. It is worth noting hence, that the traditional wisdom of a high-quality FAMT is too ambitious. The best results can be achieved either by using MAHT or HAMT.
With respect to the Arabic language, as a case study in the field of machine translation in the thesis, a number of issues related to Arabic and the Arab world are problematic and still await solutions.

Arab countries have to take seriously concerns over the future of linguistic diversity in the Information Age. Most information currently on the Internet is in English, a language that most Arab population do not know well. If this situation remains, it will create a new face of literacy in the Arab world. Those who do not have a good command of English will remain sidelined on the information highway. Many users in the Arab world today complain of the shortage of Arabic content and informational resources on the Internet.

Since Language is today at the crux of a new Arab renaissance centered on knowledge and the improvement of science and technology, linguistic research has become a critical endeavor. This requires establishing language centres, Arabicization of scientific terminology, moving forward with research into language engineering and renewal of Arabic by initiating a fresh formulation of its grammatical rules to meet the requirements of computational processing. It is also essential to consolidate and enhance glossaries of specialized terminology and thesauruses.

Unfortunately, Arab countries are still lagging behind because there is a lack of interest from the Arab financial sector in information projects, were feasibility studies are normally undertaken on a purely economic basis. Equally frustrating is the fact that there is no pan-Arab policy in Arabicization and the development of the Arabic language to better fit in the Information Age.
Access to sources of knowledge in languages other than Arabic is mainly connected to translation. In order to keep up with the pace of a world overloaded with information, and the quick development of science and technology, the Arab world must engage in a revolution in the translation industry, both human and machine. In order to achieve that, Arab countries are forced to address the challenges facing the Arabic language: There is a need to improve Arabic linguistic systems, to develop massive technological approaches in language engineering to solve problems related to Arabic language processing as a natural language, to acknowledge that information and communication technology is a tool for communicating knowledge and to take into account that the computation of the Arabic language as a basic starting point for this approach. Research and academic institutes should naturally lead in the effort to tackle both the processing and evaluation of the Arabic language in this modern age.

Arab countries, for example are developing their own models for software systems on several levels. Some of these require on-the-job-training. There is a need to train language and translation graduates in computational linguistics and to retrain engineers to develop Arabic language software. There is also a need for a basic research to build programs to handle the special characteristics of Arabic on different levels (morphology, syntax and semantics).

Arabic, as a Semitic language differs from European languages morphologically, syntactically and semantically. Most words are formed from a tri-lateral roots which falls into specific patterns; a key morphological feature. Though there has been much interest recently in handling morphologically rich inflectional languages such as Arabic, the Arabic language is somewhat difficult to deal with due to its right to left orientation and its complex morphological structure.
Because the grammatical system of the Arabic language is based on a root-and-pattern structure and considered as a root-based language, a challenging task facing research community is developing computer based algorithms and their implementations that can process common everyday use and a non-sanitized and non-novelized Arabic text.

As the corpora analysis of this thesis has demonstrated, morphological analyzers have been successful in solving morphology related issues. Syntax on the other hand, has been addressed by many researchers with only some success. What is critical to improving machine translation in Arabic lies in the fields of discourse and pragmatic.

The future of MT is bright if we remain realistic. To obtain a translation of suitable quality, hybrid and innovative approaches must be relied upon. This includes using large and comprehensive dictionaries, a wide range of data base, an advanced translation memory and syntactic and morphological analyzers which rely on unextended base of linguistic rules. In order to solve problems of text in context and fixed expressions, techniques such as parallel-corpora and statistical systems provide possible solutions for today. Future improvements in computer hardware and software and in language technology and engineering may create machine that can replace human translators. This is a dream not to be realized for years to come.

This thesis is one of a few research activities conducted in the Arab world in the field of machine translation. It is but a step with miles to go. Machine translation is a field which requires further research and development.
References


www.personal.umich.edu


Arabic Publications:


