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 EnglisWArabieiEnglish Translation and Interpreting.

By

Wesam AL-Assadi

Supervisor

Dr. Rana Raddawi

Spring 2004

College ofArts ardsciences
Department offngGsh and Translation Studies

# Declaration

I, WeSam..Fawzi..Al.Assadi., confirm that the work carried out for the purpose of this thesis and its text have been done by me. t am the sole responsible fot 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: Ha11th , 2004

Signed; [redacted]

# Table of Contents

# 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, any husband, for patiently bearing with env work dernands and Odd hours,

To my daughters Farah and Sara, thank you for allowing muni to be with her papers Instead Of mothering you,

Thanks also to my father for his encouragement and to rny mother for praying this v€01'k go to a successful conclusion,

# List of Abbreviations

| | |
|---|---|
| | Human Translation |
| | Machine Translation |
| HAMT | Human-Aided Machine Translation |
| MAHT | Machine-Aided Human Translation |
| FAMT | Fully-Automated Machine Translation |
| CL | Controlled Language |
| TM | Translation Memory |
| CL | Computational Linguistics |
| ACL | Applied Computational Linguistics |
| NLP | Natural Language Processing |
| | Language Engineering |
| | Information Technologv |
| | Language Technology |

# 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' (lumpclt, 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 or the. dominant nation...Diplomacy was in French" (Newrnark, 1988), translation in the twentieth century became a prominent factor in a world moving towards multi-linguglism, global economy and global knowledge.

In today's global economy, as more peopje increasingly need mot want to communicate with counterparts abroad, global inhabitants ar confronted with daunting cultural and language divisions. The fåct 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.

Globelism has intensified 'the demand for translation, To teach customers around the globe, businesses must affer information and instruction in the language of the target customer- If the world'snon-English speakers want to be more than passive receivers of inforrnetion, translation is needed. The impact or the Internet has been significant in recent years. The increasing multi. linguality the web constitutes additional challenges for trenslation industry. The global web can oniy 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 fot immediate translations will sunzly continue to grow rapidly and users are already seeing an accelerating growth of rea;-tirne on-line translation cm 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 urgent foe 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 barrier5 and to further the cause of international peace.

The early dreams that stimulated research and development efforts was that of a machine that wouEd 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 10 die dream if we had not given up sa soon. Negative evalwations of machine translation in the 60S were based on the argument that the understanding of text by computer wastoo difficult. The ALPAC report hy 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 Tully automatic translation between human languages. Through bitter experience, scientists have realized that they are szill far away from achieving the ambitious goal of translating unrestricted texts. Kevenheless, computational linguists have created software systems that simplify the work of human translators and clearly improve their productivity. The expectation that translation machines might repleee 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 challeJ1ges 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 cf more powerful computers. the increasing availability or large dictionary data sets and advances in some aspects of linguistic theory suggest opporn•ni€ies for research and development

Most MT research and virtually alt commercial MT activity has concentrated on the major international languages: English, French, German, Spanish, Japanese and Russian. The

Languages of the less developed couatries 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 OWTI dividing line- It can, on the one hand, lead to a wider linguistic divide between the Arabs arid 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 languegcs 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 translatiom

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 foe Arabic for a simple reason: the computation of the Arabic language, compared with English is much more complicaced 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 sinee 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 A rabic and to propose solutions-

Tlle choice to focus on MT and computational linguistics in the Arab world is ror various reasons. First, it is a domain which has not been researched enough both at the undergraduate and postgraduate levels, Therefore, students citranslation and the future Arab translators have very little idea about this fast irnproving field. Second, Arabs have yet to ride the wave of of the most profound technological phenomena in the history of mankind — the Internet, Despite the 041going 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 o? information,, Third, as for my self working in the field ofjournalistic translaqion, 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 ofthe information they are looking for,

it is obvious then that MT and language teehmiogy are most needed in the Arab world in order fot it to engage in the Information Age through translation. Machine translation and localization is a nourishing industry in the West as well as in Japan and other countries around the globe in facin! 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, Tite 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 Languagc

Processing (language Engineering). Chepter 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 ill processing within the guidelines of Computational Linguistics and Language Engineering. The chapter also includes brief overview, which covers the status of MT in the Arab leseapzh institutes It also includes a list of the pioneer Arab and international companies interested in the research and application of Arabie language technology and MT, as vve.ll as a list of some of the

mercialMT systems available in the market. Chapter four is dedicated to a corpora analysis of texts translated hy commercial MT systems from English into Arabic- The aim is to examine the standard of MT output in order to expose the »æaknesses and the strength of the Industry and to try to propose solutions. It seems that MT in the Arab world is still in its pleliminary 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. are very scarce and the number of books on this topic are but handful, so my main source of information was the Internet. My main concern was objectivity end 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 machir,e traqjslation software due to the globalization of the Internet has had little impact. The Users knowiedge of the complexities behind trans14ting remains limited and judgments are based on personal experience, For more than five decade5, people have tried to prograrn 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 chgpter aims to bring forth into the arena some of the crucial issues behind machine translation. Understanding of these particul.ar issues is the only way to move closer to the dreams ofa society no longer hildered 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 teal changes brought aboutby the information age. The globål market, the increase in intercultural contacts and the acceleration of information production, have resulted in profound changes in the vvay translators 'K.rk.

Currently, human translators must use an extensive knowledge base to ach the main task of translation — the transfer of technical and cultural information. As such, translation requires new strategies and a paradigrn shift in methodology. '[This shift must embrace practice, teaching and research," argues Austermuhl (2001, p-l).

The vor,ccpt 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, is the lingua franca of the global market economy," according to Austermuhl (2001, pj2)- Around 85 percent of international organizations use Engiish as their working language. In Europe, 99 percent of international organization: have English as one of their official languages CMai & Welch, 1999, p. 130 cited in Austermuh, 200 i), 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 Austerrnuhl, 2001)-

language now is a major factor in the debate over globalization, especialiy because Internethas made its political, cultural and economic importance universally clear. In 2001 , round 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 •+æinally published in the English language. "Concern over the future of linguistic diversity in theInformation Age is evident from the currency of such terms as 'lenguage divide', 'extinction of languages', *linguistic racism'. and qinguistic wars' (Arab Human Development Report, 2002. p.%)".

In this context, Austermuh! (2001) raises questions like: "Js English ringing the death knell for lhe rest cf the world's languages? Will the vision of monolingual world lead to the end or 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 ELF is entitled to use its own language to conduct official business within the institutions of the El.], "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'^t (Stoll, 1999, p.17, cited in Austermuhl, 2001). Since there are I I official languages and 1 10 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 teminology work)

(Austerrnuhl, 2001 0.3),

Beyond political institutions, in fact, knowledge Of foreign languages is not widely spcead in Europe, Around 28 percent of German executives have very good command of English skills. A university study conducted in 1999 inditetes that one in four German university professors aould not attend international conferences if English were the sote working language.

It is relevant to observe hete that facility with the English 1031guage is waning across the Arab qorld. "With the exception of a few univer5ity professors and educated individuals, real proficiency in English has ebbed, preventing many Arab researchers from publishing their research in internafonal scientific journa]s+, according to the Arab Human Development Report (2003). This treald 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 increases the need far translation.

The increasing cross border communications, the rapid of technical and scientiFc 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 severel 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. rmul argues that the non-English speakers are "the fastest growing groups of new Internet sets. with a rapidly growing interest in non English sites as the Net becomes genuinely nullilingual. Websites in Spanish, Portuguese, German, Japanese, Chinese and Scandinavian languages are showing the strongest growth rates" (Austermuhl, p.5).

Although around 57,4 percent cf the Intecnet users were basically English speakers in 1999, there is evidence that the number of the non-English speaking Intemet users is rising steadily as penetration rates in non-English speaking countries continue to rise. According to Computer Indust0' Almanac (cited in vvww-escawa.org., 2004), the number of' [nternet 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 Amcriea and parts of Europe. By the year 2005; the number of worldwide Internet will exceed I 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 distributi0ft of 0iiiine language population totaling ..S6i million (March2002)

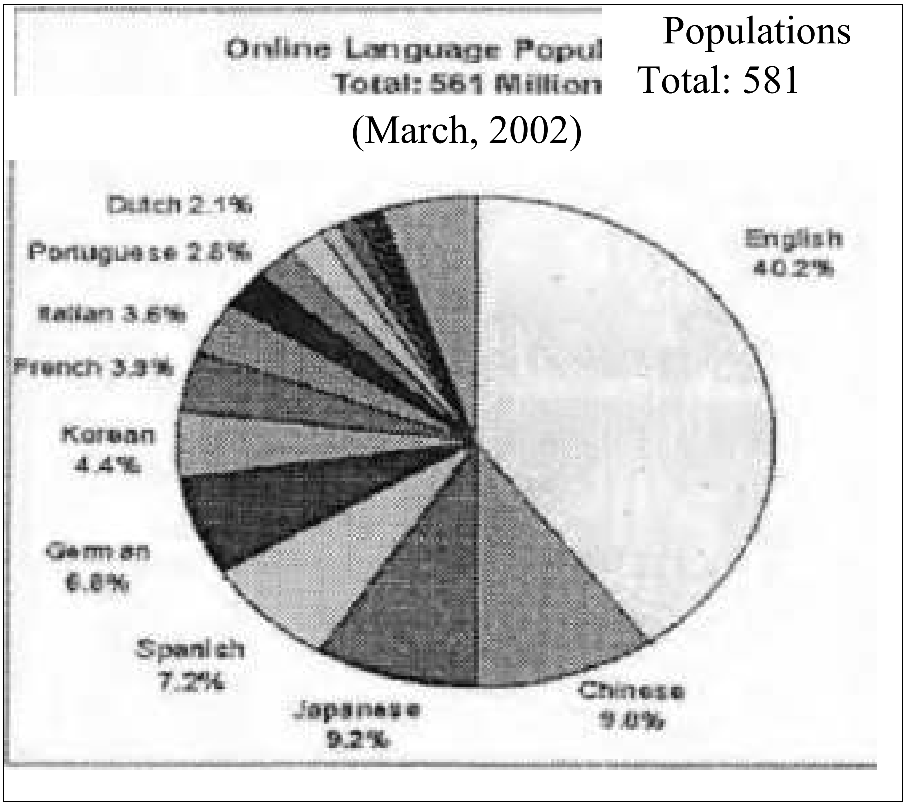


Figure Distribution of the online language population

this ill turn means that the number multilingual sites will grow and translation services and so i'iware bccoming an integral part of international communication, International Data Corporation (cited in www—l .ibm.com, retrieved on Feb 2004) estimates that the machine anslatioti software market sales record 'svere around S378 million in 2003, according to Beck, the IBM Voice Systems Director.

Not only globalization and the increasing numbers of non-English Internet users luve caused the growing demand fC[' translation but the digitization of the global economy has u 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,

A usterrnuhl (2001) argues that translation is also influenced by the enormous degree of technical speciatizgtion 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 ofknowledge accumulated during the past 25CÅ] years.
- 165,000 scientific journals are currently being published;

.20,000 scientific papers are produced every day {Mark, 1998, cited in Austermul, (2001);

- Thc amount ofdata that is circulating on the Internet an any given day is larger than 21

1 the information available throughout the nineteenth century {Der Spiegel, 1996, cited

in

Austennuhl, 2001 )

The previous figures indicate that the size of the information flood is too large for the human brain to pcocess on its own, Humans definitely need the service of electronic toois; the aid of the computer to conduct translation.

## Machine Translation

## 1.2.

1.2.1 Historical Review tine idea for a machine that would transfeF one language to another came from code breaking during the WW Il 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

theU.S. military and intelligence agencies during the 50's end 60's. By the end of the 60's the inuerest 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 biased and short sighted, hindered MT research for a decade in the LIS and in the Soviet Union and Europe as well, However, research continued in Canada, France and Germany, In '1970y MT Systems were installed fot use by United States Air Forces (USAF). [n the seme 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 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, accordieg to Hutchins.

e 1980's witnessed the development of a wide variety of MT systems -and from a number of countries, In addition 10 Systran, functioning in many pairs of languages, thn 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 Globulink. Other systems were developed by Japanese companies such as Sharp, NEC, Mitsubishi and Sanyo,

tn the 1990's, MT systems, based purely on statistical method and torpora approach, were developed. In both methods, no syntactic 0T semantic rules were used in the analysis of texts or ill the selection of lexical eqtrivalents. 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 incretse 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 is the *'application of computers to the translation or texts from one natural language into another" (Hutchins, 1986, PI).

# Machine Translation

.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 I March 2004) .Il is a multi-disciplinary field of reseanh, It incorporates ideas from linguistics, computer science, anificial intelligence, statistics, mathematics, and many other fields, If machine translation, Seasly argues, becomes accurate and efficient enough, it can bregk 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 companY5 Market, save translators' time and companies' money,

1-2.3 Different Types of Machine Translation using an appropriate terminoh)gy, there are four basic types of machine translation, Sec thc following figure adapted from Hutchins and Somers (1992, p. 148b

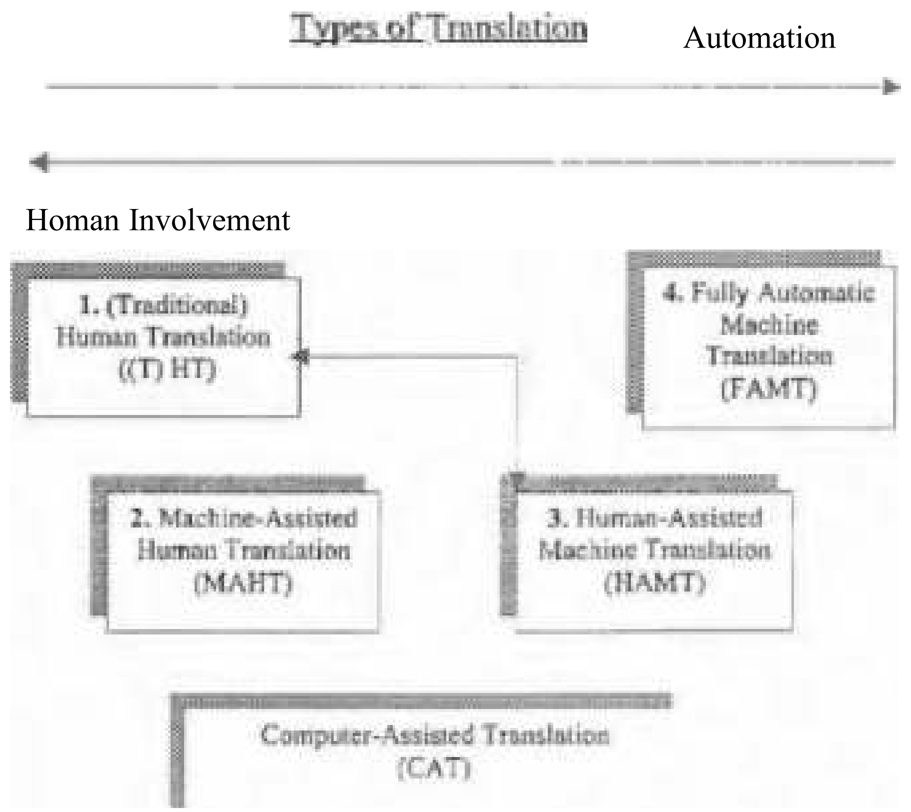

Types of Translation
Automation
Homan Involvement
1. (Traditional) Human Translation ((T) HT)
4. Fully Automatic Machine Translation (FAMT)
2. Machine-Assisted Human Translation (MAHT)
3. Human-Assisted Machine Translation (HAMT)
Computer-Assisted Translation (CAT)

Figure The four types are: Hun-Jan-Translation (HT), Fully Automated Machine Translation (FAMT), Human-Aided Machine Translation (HAM T) and Machine-Aided Human Translation (MAHT).

The first two types represent the two extremes in translation; human translation is carried out without the help of the machine, and the fillly automated machine translation. also called unassisted WIT, 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 [he MAHT is where the humans use machine's help.

## Machine Translation
## 1-2.4 Types of Demands

One can distinguish four types of machine translation demands and use of computer based translation software, according co Hutchins Cl 999):

T- use of MT for dissemination. This the first and traditional type, where the demand is for quality translations as expected from human oanslators, i.e. translations of publishable quality. However, MT systems still may produce output which must invariably be revised or 'past-edited' by human transla&ors, Ln this sense, MT systems are mostly producing *draft' translation,

2- The use or MT for assimilation. This type of MT is required for translations of lower level of quality {particularly in style). It 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 demand for translation between participants in oneto-onc communications (telephotw or written correspondence), This type is typically required for translations ot- electronic texts on the Internet* such as web pages, electronic mail and eyen electronic 'chat' lists. The need here is merely to convey basic content* hence for immediate translation regardless of qua'ity.

4- The use of MT for information access. This is the integration of translation software into s»tems 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 [illegible]&Gden the access of all EU member states to sources of information.

## 142.5 Popular Conceptions about Machine Translation

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

l) MT is a total waste of money & time The quality of output is generalty very poor. The traditional anecdote here is that of the Russian Mt system that translated The spirit is willing, butlheflesh is weak into the Russian equivalent The vodka is good hul rhe sreak is usy.

2) MT will break all iangusge barriers in the global stage. Injust a few years time the output of machines will be as good as humans' output.

Although machine translation may not provide a complete so.lution to the problems of translation due to the unique and complex nature of natural language, it can be an efficient ml in translation of text, at least restricted knowledge domain. Ln 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 gf MT requires certain tools to improve

•the quality, when needed.

chapter four, some samp[c texts will be translated by using MT software systems to show how helpful machine translation can be and where it needs improvement.

## 12.6 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 where necessary, editing the ST before translating it, so as any strings of text that an MT system wil! have problems with, are highlighted end 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.

- AControlled Language, by definitions is a subset of a natural language whose grammar and dictionaries have been restricted to reduce or eliminate ambiguity and complexities in texts mitten in that CL.

- Translation Memory Systems arc basica)ly the building of a translation dåtabase 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 y.reaus, Many translators, translation offices and companies build their own in• house =tionaries in order to ensure consistent usage. On the other hand, there ale now numerous tilingual 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 tion 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 TU. Hutchins (19S6s p.34) argues that:

> The basic assumption that the vocabulary and syntax of SL texts should be analyzed no mote 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 includedw was testricted to the use of features such as 'male', $^{V}$concrete' $^{V}$liquid% for resolving collocational ambiguity.

In the $^{i}$ second MT generation', however the direct approach was abandoned and replaced by inier-lingtita\ or rransfer approaches where ttmslation is indirect via inet•mediary 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.linguai' and *transfer' systems, the analysis of SL texts is quite independent of the TL As Hutchins (ibid) states, "The systems are not '-hetefore designed for translation only between two specific languages, but can in principle be udopted 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 honiversal' representations, After the shortcoming appeared ivi a number of MT systems, MT researchers were convinced that "the appqoach to linguistic modeling adOpted ov assumed hy the 'pure' linguistic theorists is not appropriate," (Hutchins, 1986. p, 12),

Inorder to achieve praeticgl objectives of producing quick translations of technical documents, it is time to take a more pragrnetic 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 translutions,, In recent years, there has thus been a number of 'interactive' MT systems under developmentInteractive systems are most attractive where there is a need for simultaneous translation of a sinelc 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 end semantic links between sentences — those feature which — take a sequence of sentences into a cohesive whole have been neglected (Haliday and Hassan,

A subject of current interest in machine translation (in the United States in particular), according to Foster et al, is the r4Pid development of systems for novel language pairs, At least one of the languages in question ig taken to be previously unknown to system deve!opers "'who must either acquire the necessary knowledge and technology ot devise methods that will mitigate the effects of their absences... this is viewed as a counter to excessive reliance on exceptionally iargc and 'clean' parallel corpora", argue Foster et al (2003).

Researchers today are competing to improve thc quality and accuracy of the translations. A'though statistieal.based translations are not 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 Ozbek, Bangali, Nepali and others. The ambition is to develop systems for as many as 100 languages within rew 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 further develop MT systems, it is essential to see how far the theories of computational linguistics serve the field of MTL

# Chapter Two

## Approaches to Machine Translation

## Theories & Applications

The linguist's contribution to the modeling or translation process has been always recognized in zz•ysiation 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 Computalional

Ereujsr}cs.

## 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 oi success eased on whether or not the *Stem is an accurate model of the human brain or simply an •elegant" theory. The search for a universal grammar for al] 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 b2sically 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 languages. For every need, a fitting solution must be found.

Co—putational linguistics (CL) is "a discipline between linguistics and computer science which s concerned with the computational aspects of the human language faculty", according to L süoreit (retrieved on March 2004), It belongs to the cognitive sciences and overlaps with the field of Artificial Intelligence (Al), a branch of computer science aiming at "computational models of human cognitior•p.

CL. according to Thompson (1 985), is thit 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 computø\ionally". In other words, CL concentrates '"on studying natural languages, just as traditional linguistics does, but using computers as to model (and, sometimes, verify or falsify) fragments of linguistic theories deemed of particulat 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 Uszkoteit (1985). Today these theories have reached a degree of cornpiexity that can only be managed by computers- CL develops formal models simulating aspects or lenguage features and implements them as computer programs,

In addition to linguistic theories, findings from cognitive psychologv play a major role in simulating linguistic competence,

Applied CL focuses on the practical outcome Of modeling human language use. The methods, hniques,tools, and applications in this area are often subsumed undet che term Language E%ineering or (Human) Language Technology,

the goål of CL to create software products that have some lutowledge of human language. They are urgently needed for improving human machine interaction since the obstacle in the interaction between human and computer is a communication problem, "Today computers so not understand our leneuage but computer languages are difficult to learn and do not cotrespond to the structure of human thought," according to Uszkoreit. Even if the languages ahe machine understands and its domain of discourse are very restricted, the use or 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. Multilinguality obviously presents problems fot any theory that assumes a text to be embedded in a petticular language.

The comparative work carried out by nineteenth centUt•y grammarians was concerned with establishing an explanatory basis for the relationship belween languages and groupsof languages primarily in terms of a cornrnan ancestor. The comparative grammar, in contrast, is "concerned with R theory in grammar that is postulated to be an innate component of human brain", according to Nabi[ Ali (retrieved on Jan 2004)- In this way, the theory of grammar is a theory of human language and hence establishes the relBtionship among all languages, not just thoseIhat happen to be related by historical "accident" {for instance, via common ancestry),

Che characteristic of modern linguistics has been the attention given t? the formalization of descriptions of linguistic systems. While vigour and precession have been a feature of the •vitings of linguistics since the late nineteenth century (e.g. The Neogrammarians), it is primarily the work Chomsky (1957, 1965) which has placed uJTiversal grarmnae at the centre Ofthcoretica} inguistics, 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 stiff uncertain what kind of grammar is appropriate for natural language and what the general characteristics of the formal model should he. In consequence, much of the theorizing in linguistics about the (orm of grammars and about the formal treatment of particular linguistic phenomena (ease rclatians, semantic features, transformational constraints, pronominalization, passiviætion) is carried out in a vacuum Wilh no direct contact with real linguistie data.

Ingeneral, linguists have tended to ignol± problems of translation, "The theory of translation is one of the least developed areas of modem linguisLics", Hutchins (1979). The common attitude can probably be summarized as: •'we cannot: yet describe linguistic process involving one language only, let alone attempt to de5cribe what goes on in translation" he argues. Why then shouid 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 trnslation ofanother text

another language and can be evaluated by independent judges, It provides a clear test of the rightness or wrongness of z proposed system, since the output in a second language can be .sgssed by people unfamiliar with the internal formalism and methods (Wilks 1975, cited in Hutchins, 1986 j. The evaluation of transl.ations has its problems, but in principle it can be objective, e.g. by ahserving whether the users of a manual produced by MT can under-stand 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 [anguage,

.4ccceding to Hutchins, there are probably many reasons why linguists have generally been unwilling to be associaæd 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 e mistaken conception of the real aims of machine transletion". The primary stimulus for MT research has always been the urgent needs for scientists, engineers, technologists, economistsv administrators, etc. to cope with 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 —as-uages. It includes the development of algorithms for parsing, generation, and acquisition of —guisric 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 softwaje architectures for •tious NLP tasks; end consideration of the tvpes of non-linguistic knowledge that impinge on NLP. It is a fairly abstract area of study aid it is not one that mzkes particular commitments to the study of the human mind, nor indeed does it make particular commitments to producing useful artifacts, according to Uszkoreit (2004).

unguage 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 or 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 zlzematives*. The main al ternatives include partially oc 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 appareJ1t to all users or potential purchasers of systems. As a consequence they are sometimes unable to distinguish Sett€eerl the failings of particular systems and the problems which even the best system would have.

Transia,tion is essentially a problem-solving activity, choices hayc to be made continually, The asstrrnpGon in MT systems, whether fully or partially automaticm is that there ate 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 blems 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 tuations, what kind of information is appropriate in particuiar 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 he outlined.

## 2.5. I Types of Linguistics and Computational Complexities

There are many challenges to Machine Translatiom Some of them are:

(i)the use of other specific words in the seme phrase or sentence iiii

the use of morphological information

4iii) the use of information about syntactic functions and
relations (iv)the use of semantic Features and relations 6 ) the
use of knowledge about the subject domain
(v i) 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 orword Oider, Three types of plublern will be discussed; compound nouns, idioms. and metaphors.

All translators are familiar with the need to treat compounds as units to be trans!ated. In many eases an attempt to translate each component of a compound noun would lead to ridiculous results:English •eggplant' is not ' in Arabic, but Many potential problems of homonym can be averted by the entry ot- the relevant words in combination with others in dic€ionaries,

The word light for example, can modify another noun in at least three different senses: an adjective 'not heavy', an adjective dark' and a noun ;tumineseence 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 mBke use of this fact, Thus many MT systems include entries for compounds such as a light ship and a lighf bulb; and indicate *Qtly the target language equivalent (French ampoule, Germail Gluhbirne). In this the system c.an avoid a perhaps lengthy disambiguation process to determine which of the two stses of bulb is intended ('plant bulb' or 'pear-shaped glass') and combination with which of the three senses of light; a process which will have to be done every time the compound ig encountered.

S.nrne would argue that the mast difficult area for MT must be the apparently unclassifiable variety of idiomatic expressions. [1 is a \iew Which has support in the stories of early MT s:.S1ems which translated "oaf ofsigh! afryihd "as *invisible idiot',

The perceived difficulty of Idioms is Chat the individual words take on meanings and connotations, Which they do not have, in their literal usages, However: according to Hutchins, it is preccsely because most idioms are relatively fixed expressions consisting of the same words in the same sequenee, 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 common thread of similar formations so that even if a metaphorical usage is not recorded in the dictionary, it may bc possible produce a 'literal' translation which has the same metaphorical impaet, However, it would be a weakness in any MT system if it did not account easily for many metaphors, which have become standard expressions ofthe language,

## 2.5.1.2 Morphological Analysis

One of the most straightforward operations of any MT system should be the identification and .•eneration of morphological variants of nouns and verbs. There are basically two types of morphology in inflectional morphology, as illustrated by the familiar verb end noun paradigms (Arabic 1.5-1-4 يأكلان، يأكلن، يأكلون etc.), and derivational morphology'E which is concerned with the formation of nouns front verb bases, verbs from noun forms, adjectives from nouns, and so fortll (e.g. الوطن، يوطن وطنية وطن etc; and equivalents in other languages.

It should be stressed that any MT system should as a minimum he capable of recognizing morphological forms and cf generating them correctly, However, the alignment of equivalences between the verb forms among languages is another matter, particularly when modal forms are involved 0'igh1, devoir, faj{oir, mogcn, dur/em, etc))

In general, a MT system which cannot go beyond morpho}ogical 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 Ofeny treatment of word order will give poor results.

## 2.5. L3 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 thc 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 *jecmes fille.s son/ venues, die meisreyt Frauen Sind nick/ gekomrnen. Of course there are complexities in the syntactic analysis. [n 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 Enghsh, the major problem is the categorical ambiguity or 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 rnarke/ car. 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 marker as a prepositional phrase, and rose quickiy as a verb phrase,

In addition to syntactic problems, this section of sentå]ltic rotes and features also wrnps up difficulties in MT,

### 2.5.1.4 Semautic roles and features

The recognition of implicit relations rnay well require access to semantic information- It is common to identify two types: semantic roles and semantic features, By the semarüie roles in a structure is meant the specific relationships of nominaj elernents {entities) to verbal elements iactioms or states): a petticular noun may be the 'agent' of an action, another may be the instrument$^b$ (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 Developers of MT systems are usually obliged to draw up their own list. However, the pTincipal difficulty is the identification of roles, Hutchins argues. In English, the main indicators are the propositions, but these can be ambiguous as to role expressed; with can indicate instrument, manner or context:

- the bottle was opened with a corkscrcw -

the bottle was opened with difficulty - the

bottle opened with the meal

Semantic features refer to labels such as 'human', 'animate', •liquid', 'young', etc, assigned to lexical elements. They can used either in conjunction with semantic roles or independently, For examplet 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 lunge hat die Banane gegessen
- The monkey ate the banana→ Der Affe hat die Banane gefiessen

Such features have to be assigned to ail 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 complied their own lists, some are minimal and rigidly controlled and others are extensive or not applied eonsistently.

## 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:

(l) old men and women les vieux et les vicilles or.

les vicux et les femmes

(2) pregnant women and children— des femmes enceintes et des enfants

not; des femmes et des enfants enceintes

In (l) we have no idea, out of context, whether '401&' applies to tXJth men and women or only to men- But in (2) we do know that "pregnant" cannot apply to children; it is part of our knoWIedge about women. This knowledge needs to be incorporated in the MT dictionary in someway, prohubly by limiting the use of "pregnant" to nouns with the semantic features 'female' end 'mature'.

Similar problems arise with relative clauses;

- Peter mentioned the book I sent to Maw

  → mentioned the book (which J to Mary)

- mentioned (to Mary) the book (which I sent)

ne '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 contexteven 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, transletion goes beyond the familiar linguistic information: morphology, syntax and 5emantics,

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 in formation which be refereed to during the translation process, It is the approach cotnmotily referred to as that ofAt'!ifieiat Inrelligeme 040. For exemple, given a sentence such

the following occurring in documents relating to computer hardware

- Remove the tape from the disk drive

The word tape can po{entially refer to a 'magnetie tape" or an $^{V}$adhesive tape'. An Al-based system would check in its knowledge bank svhiieh is most plausible in this context, i.e. it would seek to answer the question whether tnagnetie tapes can be removed from disk deives, or whether disk drives can contain or have as parts items which are magnetic tapes, If not, then it may cheek whether 'adhesive tape' is plausible, i.e, whether disk drives ane 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 zougl-•.. Coverage of any documents other than those within a narrow subject range would clearly require databases o? massive proportions. While the computer hatdware and the .c,.r-nputer software for fast access may well bath be already available, the databases are not. These would demand many years of difficult and complex work by many researchersTherefore, it is not surprising that MT systems are based on well-known teehniqwes 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 eun be guilty of this fault as well — although Newrnark (1988) considers literalness to be desirable in literary and authoritative texts, as long as the result is in the appropriate style, Hgwever, the aim in technical translation is generaRly to produce texts which read as if they wcre 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 md 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 exhaustcd al] 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 systemitself 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 increzsing number of MT users, who require texts to conform to certain restrictions of vocabulary and syntax: certain words are to be used in one mean ing 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 syltactic analysis,, The use of semantic features and roles is also well researched and reliable, With these techniques it is possible to deal with 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 greeter difficulty. Some may not occur often i" cettåin text types and some may not be eritiea] for certain users (i.e. they can be handled easily 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 being developed to improve the output efficiency ot-machine translation systems, Various countries around the globe are competing to improve MT systems that serve their commercial and poiitical interests,

## 2.6 Machine Translation in Use

In this section provides brief overview of MT statL15 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 coveted. MT status in the Arab world is scanned in detail in chapter three,

### 2.6. I 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. [n Europe, there is a need to address alllanguages of European including the language of the new members in the ELI from Eastern and Central Europe. In the USA, they feel that they have strong strategic disadvantage: every onc 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;

:.6.1.1 MT in the United States t; is warth 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 Computingv Information and Communication (CIC) Program Bhich started in with budget of USS I billion pet year, according to Marrai (2004), Several parties receive support from this budget, amongst which are: the National Science Foundation (NSF), uhich supports basic research in Foundation which encourages the basic research in Speech and Natural Language; the Defense Advanced Research Agency (DARPA), which carries out cote technology development; several national agencies including the CIA, FBI, uS Air Force, Dept of Ena•gy, National Security Agency (NSA) and others, which develop applicatiom

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

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 LIS currently leads Japan in technological diversity, that is the variety of approaches to MT, and linguistic diversity, that is the number cf 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 tnultilingual companies to mention are; Ericson, where the Logos system is providing 10% or tnnslation needs for producing manuals and dacumentat.ion in French. German and Spanish); SAP, using METAL for German-English translation and Logos for English-French (totaling some 8 mi]li0J1 words per year)'9 and Siemens providing a service based on METAL, The European Commission, the use of Syscran 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, snd sales have been lower in Europe. However, there are notable European products: the Comp:endium and Tl systems {Sail Labs), Persoml Translator PT (linguatec), the iTranslator series (originally Lernout & Hauspie, now Mendez), the Reverso systems (Softissirno), the range of PrcfMT systems (for RW55ian to/from English and German); and the PARS systems ror Russian and Ukrainian to and from English. According to Hutchins (2004), most of these systems are availab)e in different versions for large enterprises, for independent professional translators, and for occasional (home) use, e.g. for transl.ating 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

InJapan, machine translation is viewed as an impotunt strategie 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 s:.stems commercially. There is also an active MT and natural language processing research community at some of the major universities and gmemment/industrial bodies.

"It is no surprise to find that half of the world's MT research is round on that densely populated archipelago (}apan)", states Brace (2004), .lapan'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 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 systetns in practice do vary from the nearly direct Penses system of Oki to sophistisemantically rich systems like Toshiba's Astransac and Fujistsu'5 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 foc 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 IS constitutional languages there. Currently, this translation is essentially manval, 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 formatst parliamentary questions and answers, pharmaceutical information and legal terminology and judgments (Srikanth et al,

"InIndia's multi-linguistic lathdscape, where the need to facilitate smooth communication between the Centre and the staaes is vital for good governance, machine translation offers a great solution to this problem "$^{i}$, argues Srikznth (ibid). The social or political importance of MT arises from the socio-political importance of translation in countries where 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 inf&mation more effectively,

owever,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 11T Kanpur, University of Hyderabad, NCST Mumbai and CDAC Pune. The Technology

Development in Indian Languages (TOIL), an initiative of the Department of IT, Ministry of Communications and Information Technology* Government of India, has played an instrumental vole by funding these projects, Since the mid to late 90's, a few more projects have been initiated at 11T Bombay, 1 15T Hyderabad, ALI-KBC Center Chennai and Sadavpur 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-3ine 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-languege information where top quality output is not essential. Many PC-based systems are marketed foe che trznslation of Web pages and of electronic mail, and there is great and increasing usage of MT services {often free), such as the well-known •Babelfislt' on AltaVista.. At the same time, the Intemet 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 velue' by human post-editing

Finally, the Internet has also demonstrated an urgent need to replace the existing systems, developed ror well-written scientific and technical documents and assuming human post-editing, by systems and translation aids which are developed specifically to dca] with the kind of colloquial (often il l forrned 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 lhe voluminous data available on the Internet itself to form the basis of future systems for this application", argues Butchilts (2004),

Inshort, 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 theaper translation in addition assisting ltL1jnato translators improving their productivity and efficiency in translation,

Arab countries are na 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 internetional websites are overloaded with infomution wyitten 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 af translating technical and scientific books which there is a demand in the Arab world.

The next two chapters are dedicated to MT in the Arab worlds 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 exttetneLy scarce and is not keeping pace with the globalknowledge explosion.

According to the Arab Human Development Report (2002) issued by the United Nations Development Ptograrn {'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 ccntury 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 oftranslation in the Medieval Arab World, At that time, trans;ations according to Faiq (2000}, piayed a vital role in the establishment of Arab-Islamic eulturai 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 becamea 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 *'is era in history, Because Medieval Arab translators were under pressure, they adopted three main strategies of translation; tran51iteration, 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 short time and gist translation was used at the point when the need for more trans'ation diminished when Arab scholars started writing and publishing their own research. Translators then worked with linguists arid grammarians to eoin 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 eiectronic too:s 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 globalizcd world, If the outcome of MT translation is or.acceptable (again transiiteratian, 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 systetns,

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.

In-lis chapter covers issues or interest regarding Ara$ic machine translationr Also included in the chapter ave: the crisis of Arabic language, computational prcve.55ing of Arabics theoretical zpproacbes to Arabic. Arabic as a Natural Language and language engineering and research programs proposillg solutions to the complexities of Arabic as a Natural Language, A survey is given to the Arab research instiwtes incerested in machine translation; reseamh and applications and pioneering companies in the field and finally a list is provided for the Arabic commercial machine translation software systemö

## 3.1 The Crisis of Arabic Language

Language is today a recurring topic in thc 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 technologv is at a watershed. It can maintain linguistic diversity, a choice that entails difficult communication and hinders flow of information 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 langu&ge, 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 or tmZrmation technology presents a real cha21enge to the Arabic language today,

According to the Human Development Report 2003, issued by the United Nations Development Program(UNDP), central to the Arabic ianguage crisis are the following: filSt, there is a marked absence of lipguistic 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 Arabieization of the sciences and various other discipl%ies 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 fitjrn 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 Arghic. Sixth. Arabic electronic publication is weakened by the scarcity of advanced Arab software. Finally, the Arabic language continues to suffer from the dupl 'cation of research and development projects and the absence of co-ordination among them, "conflicting diagnoses ofthe ills afflicting the language, and thc conspicuous absence of a clear vision of linguistic 'eform"(HDR, p. 123).

## 3.2Complexities Of Arabic Processing as a Natural Language

Avgbic, as a Semitic language, differs from European languages morphologic*lly, syntactically and semantically, There has been much interest recently in the handling of morphologically rich inflectional languages such as Arabic from a computational perspective,

Severalworkshops in recent years (both regional ami affiliated with international conference) haveaddressed 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 Arabie arid other Semitic languages requires treatment of unique linguistic characteristics, some of are common to all Semantic languages; others are specific to each of these individual languages,

Natural Language Processing is needed because around 75% of all information is textual. In oeder 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 pure charactet 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 had to draw on its resources mnd techniques". Shortly after starting their research, Arab researchers discovered that these techhtques were not suitable for Arabic, This is simply due to the fact, according to Ali, that Arabic as compared to English is "much mote comp)ex at almost atl linguistic levels, with phonology as the sole exception" (ibid).

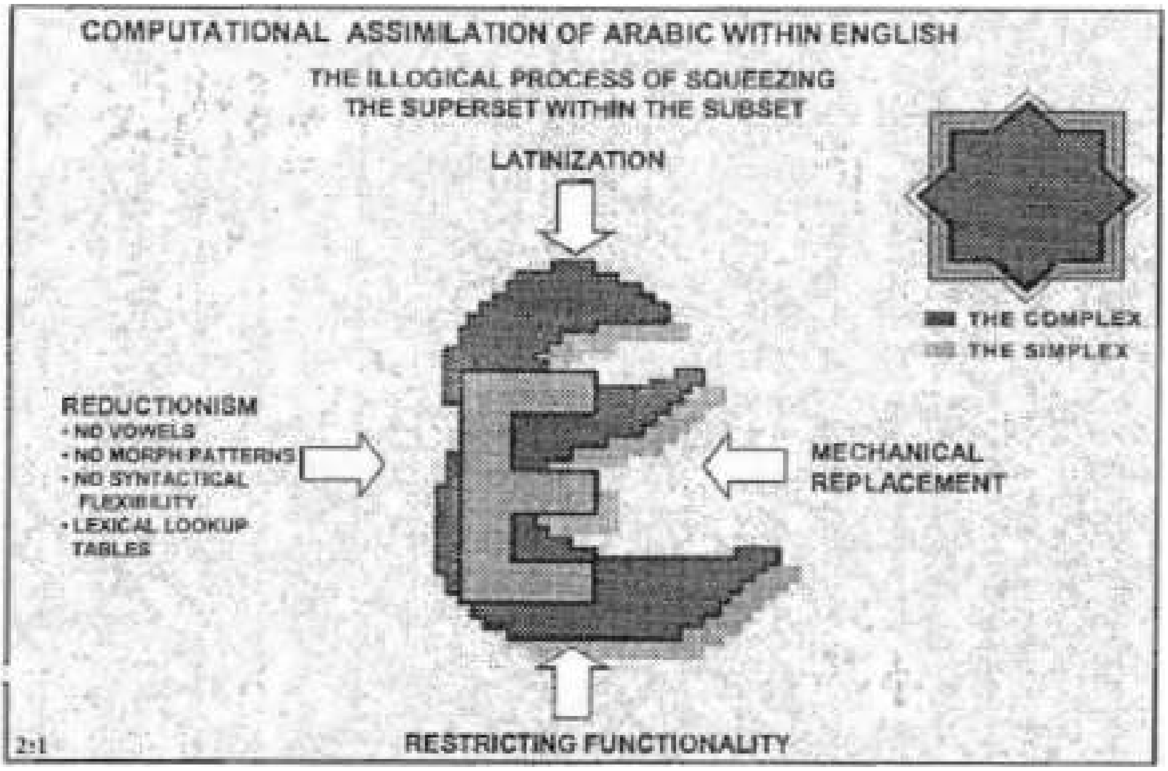

COMPUTATIONAL ASSIMILATION OF ARABIC WITHIN ENGLISH
THE ILLOGICAL PROCESS OF SQUEEZING
THE SUPERSET WITHIN THE SUBSET
LATINIZATION
THE COMPLEX
THE SIMPLEX
REDUCTIONISM
• NO VOWELS
• NO MORPH PATTERNS
• NO SYNTACTICAL FLEXIBILITY
• LEXICAL LOOKUP TABLES
MECHANICAL REPLACEMENT
RESTRICTING FUNCTIONALITY
2:1

Figure Prepered by Ali (E.SC.WAL 2003) to demonstrate Arabic Assimilation within Enish.

Since objective of t•esearch into natural language processing is to make computers deal • •intelligently with llje diversity and complexity and variation of human natural languages"i according to Yaseen, et al (2003), Arabic language processing is considered ane of the most difficult among ihe Semitic and non-Semitic languages due to Ibe complexity of its automatic processing, Research in Arabic NLP is very rieh in areas such as morphology, moderate in syntax analysis and stifl nor 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 or 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 semantie processing is required,

Chalabi, Head of Sakhr Research Centre in Egypt, told the reseavcher via telephone and e-mail that since the Arabic htnguage is campo%tionally one order of magnitude more complex than its Latin counterparts, it is unrealistic to impon 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 Atabic NLP components which took more than 15 years, with an average team of flfty linguists, engineers end 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 correspollding morphological analyzer for English.

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

- On the character level:

- On the word level;

a) Character context sensitivity

b) Overlapping

b) Different writings for some characters (Alec,

Maksoora and Hamza).

e) Diacritics and points

a) Highly inflectional language

- On the syntax level:

a) laek of diacritics in written text.

b) Free word order

e) Rare use of punctuation

Chalabi argued that native solutions specifically built to tackle the Arabic language have proven to be 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:

- Pan of speech disarnbiguation
- Word sense disambiguation
- Pronominal reference solution
- Elliptic personal pronouns deuction
- Named entity detection

For Ali, (1994. p„3SS), the complexity of Arabic at the character levei lies in the cursive shape and concatenation or Arabic letters, and above ail 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( & ) At the word level, the morphoiogy 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 zognized for its wide syntactic transformation. mechanisms like anaphora and cataphora ex:.

بلغ الفطام لنا رضيع instead of بلغ لنا رضيع الفطام14, substitution and ellipsis (such as using the

subject instead of the verb), ex: ضاربا instead of eW $1 According to Ali, in order to process Eng)ish syntax computationally, around a thousand arithmetical rules were used, whereas more than twelve thousand rules were used for syntactic Arabic computationai ?åocessir.g. Ali argues that, in essence, written Arabie 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 unveis is intermixed with other types of ambiguities, mainly those associated with word sense, pan of speech end syntactical structure. Ali provides an example to explain such a problem;

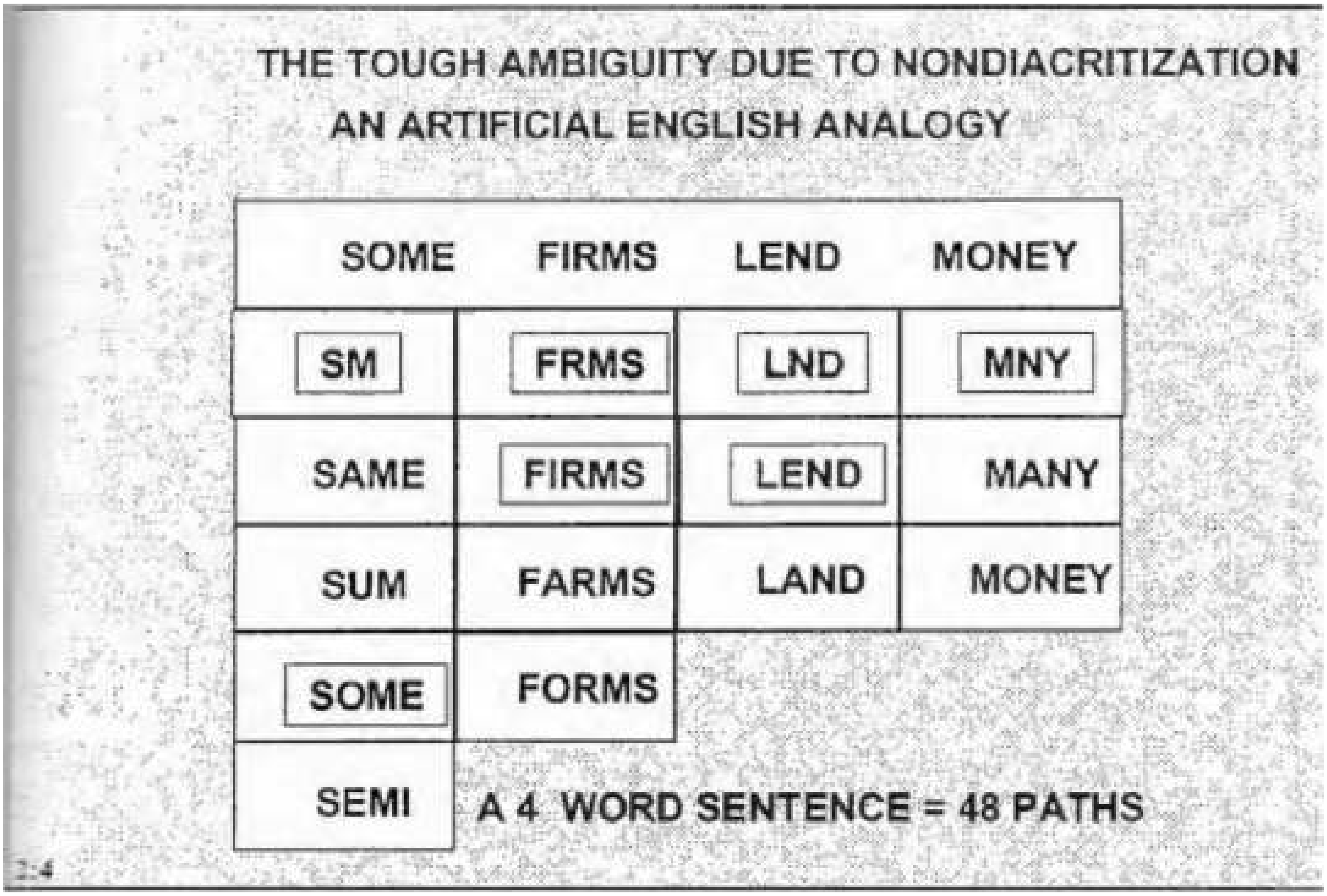
Assumed sentence: 'some firms lend money'.

The sentence as would he written in the Arabic fashion;

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

The result as it 4ppears 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 vowelizet became mandatory as prerequisite for Arabic computation. To solve this problem, Ali has developed an order to disambiguate the unvowelited text, as well as to substitute the missing vowels. This required the achievement of the three main computational linguistic tasks; I) the development af an Arabic parser, 2) the development of a lexical.

semantic processor and 3) the development of an automatic generator of the vowe lized text.



THE TOUGH AMBIGUITY DUE TO NONDIACRITIZATION

AN ARTIFICIAL ENGLISH ANALOGY

| SOME | FIRMS | LEND | MONEY |
|---|---|---|---|
| SM | FRMS | LND | MNY |
| SAME | FIRMS | LEND | MANY |
| SUM | FARMS | LAND | MONEY |
| SOME | FORMS | | |
| SEMI | | | |

A 4 WORD SENTENCE = 48 PATHS

2:4

Figure Assimilation by Ali (2003) to demonstrate ambiguity in Arabic due to nondiacritization:

Since parsing techniques developed for English have been proven inadequate for the Arabic language, both in fill',ction and performance, a parsing system based on a multi level grammar was deve loped and implemented* according to Ali, "This system is capable oihmdling 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.

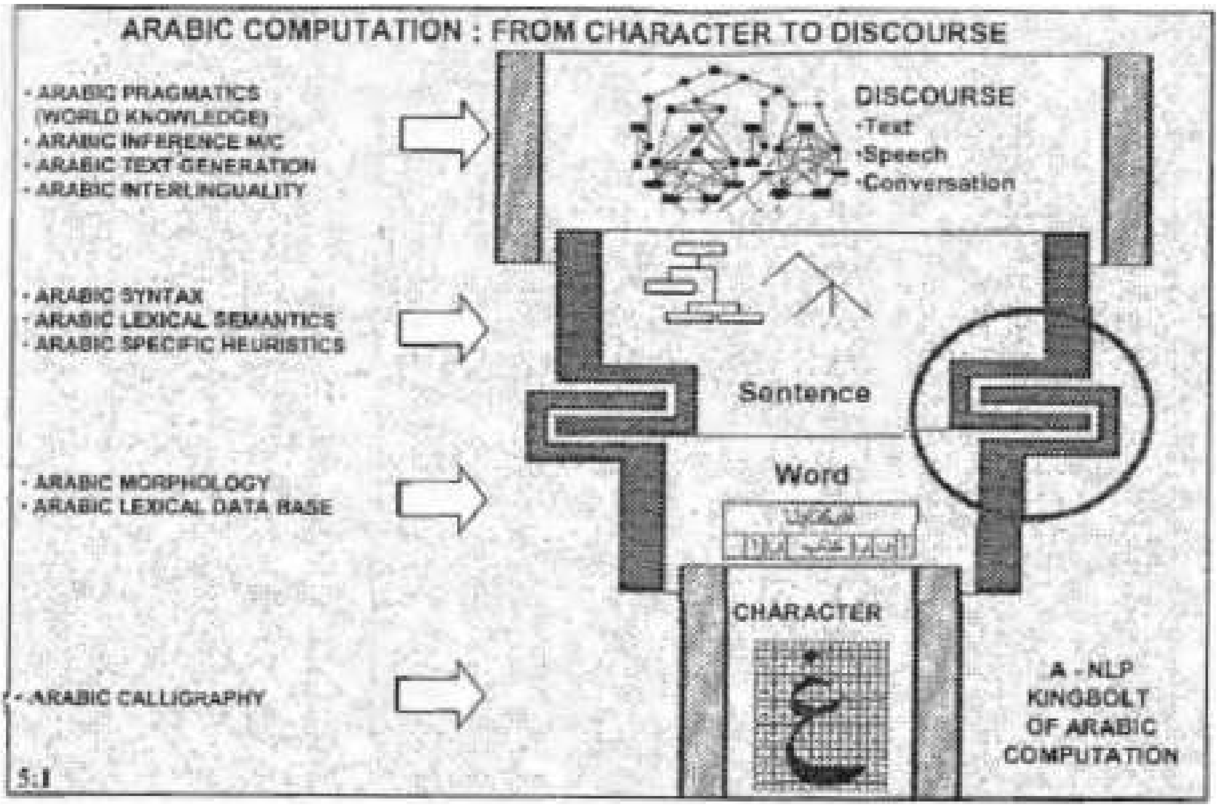


Figure The assimi!ation of Arabic computation designed hy Ali (20031

Inorder to solve the linguistic complexities of Arabic language, especially in regard to its con•puta.tional processing as a NLP. [he next part of the chapter wül 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 approaches in machine processing of Arabic. One set of approaches can be qualified as •particularist' because they '"emphasize the linguistic idiosyrasies 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 (rett'ieved 0"' 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. Guider argues that the 'particularist' approaches are cotcecned mainly with the morphological and semantic aspects Of the Arabic language, while the [i]universalist' approaches emphasize the syntactic aspects of the linguistic system,

However, Hannachy a specialist in 'the field of Computational Linguistics, to]d 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 la.nguage technolog in Arabic that are efficient and feasible.

## 3.4 Arabic Language Engineering

Most mainstream language engineering techniques have been developed Western European languages. These techniques, though superficially qune distinct, are built according to formal algorithms that machine can 'realize',

In an interview with Al-Khaleej Daily on 12 Ja;wary (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 (ew complexities, (t is basically built specific roots and patterns for verb forms and for nouns and adjectives derived from verbs. Roots constitute the basic 5ke}etoo of words in Arabic, whereas patterns constitute their overall structure. According to Hannach (2004), this mathematical architecture of the Arabic language makes it more 'füsional% in contrast to some other languages which are $^{i}$amxational$^{V}$, In order to develop promising Arabic information and communication techno:ogy applications nachine 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 wellas 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 othet languages. Implementing this "knowledge based" translation process requires cremendous 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 Ambie speakers access to the 'fruge Latin accumulation knowledge over the web and across the Internet in Arabic".

Research activities, both linguistie and technical, are crucial for the levelopment 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 shoncomings in machine translation.

## 3,541 Models of Research Projects

This section includes two research studies,' *Finite-State* Morphological Analysis a,'ld Generation ef Arabic at Xerox conducted at Research Department in Xerox and Towards Undersianding .4rabic: Logical Approach for Semanfics conducted hy Haddad and Yaseen,, The two papers aim at demonstrating possible solutions to overcome the complexities facing the automatic processing of the Arabic language a natural language, Each research paper covers different aspects of complexities; the first one covers the morpho-syntactic eomplexities whereas the second papet exarnines the semantic problems.

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

Xerox Research Centre in Eutope has demqe]oped g 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 ferm as a '"black box module that accepts words and outputs morphological analyses" (2001b

In computer analysis of Arabic, or of any other language, the input words are in digital fount with the characters in standard encodings like ASM088$9-6 and Unicode, As (or the content cf the morphologicål analyses, they will always be somewhat theoretical and applicationdepetldeitt. In short, a morphologieal analyzer should separate and identify the component morphemes of the input word, labeling them somehow with svffeient 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 serztences) 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. Bot if the leng'*age contains infinite number of sentences (us almost all languages do), the same sort of "recursive ot iterative description must be provided to characterize the sentences".

Aecording to geesely (2001), lexicons and morphotactic information are encoded in the lexc language, which is a kind of right-recursive phrase-structu1Z grammar, and are compiled into finite-state transducers, Finite-State Transducers (FSTs) are data structures that encode regular relations, which are mapping between two languages. For human convenience, a finitestate 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 fie other language, Beesely explains.

The upper-side or analysis strings of an FST compiled from the lexc description eonsigts of underlying morphemes and multi-character-symbol t;gs like; —Noun, +Vetb, +Adj[jectiveJ, •+ConjCunctiom, +VPref (verbal prefix), 4Masclcuiinel, +Femlininel, +Sig[ular[, +Plur[lal], etc that idemiö$ the morphemes. These tags have multi-character print names that are chosen and spelled according to the caste and needs of the cieve;opers, 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 vomputer analysis of Arabic or or any other language, the input words are in digital form, •.vith the characters in standard encodings like ASM088S9-6 and Unicode,' As for the content of the morphological analysis, they will always be '"theory-application dependent"- According to Beese)y (200)), the morphological analyzer in che broadest terms, should seperate and identity 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 eonjuncticns (wa.) and (få.). 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 aigorithms, and the results are stored and run as finite-state transducers. The finite-state approach to morphology, using a variety of software imptementaåons, 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 axe rprescnted as a concatenation of a root and a pattern as shown in Figure(6):

Upper: Iktb&CaCaCl+Verb+FormI+Perf+Act+at+3P+Fem+Sg Lower:
Iktb&CaCaCl at

Upper:[bny&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg
Lower: Ibny&CaCaC1 ut

Upper: Eqwl&CaCuCl+Verb+FormI+Perf+Act4a43P+Mase+Sg
Lower: [qwl&CaCuCl a

Figure 6: Three pairs of strings in the lexicon EST were compiled from the lexc description, These examp'es correspond to the words that wi[] eventually be katabåt (

the verbal from I perfect active pattern CaCaC or CaCrC and the third-person femin ine singular suffix at or the third-person masculine singular SUff1X u. The square brackets are used for convenience to delimit the stem components from the test of the word, and the ampersand serves here as just a delimiter between the coot and the patterns which are simply concatenated together. 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 mgpped via finite-state algorithms arid altemation rules into properly spelled surface strings. The first step in the modification of such strings, according to Beesley, is to interdigitåte the roots and patterns to form sterns, 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 perfonrjed much more efficiently by a finite-state algorithm called MERGE.

The application of the merge algorithm to the 10',ver side of the relation is performed by the COMPILE-REPLACE algorithm and the result is shown in Figure(7):

Upper: Iktb&CaCaCl+Verb+FormI+Perf+Act+at+3P+Fem+Sg
Lower: katab at

Upper: Ibny&CaCaCl+Verb+FormI+Perf+Act+at+3P+Fem+Sg Lower.
banayat

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

Upper:
Lower: qawul

Figure 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 (L.L.') and banayar and qawulo, which involve weak radicals and await the application of alternation rules to map them into their final orthographical forms (بنت) and respectively. Note that the upper-side strings have not been modified.

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

Figure 8', Creation Of Transducer, The represents the composition operation Nat surprisingly, to anyone who has studied Arabic, the rules controlling the realization of w, y and the harnza (the glottal stop) are particularly complicated. In the examples shown here. katabait is finished and can be displayed as the underlying final y radical of banayar disappears on the surface, leaving banar ( and the underlying medial radical w of qqwula disappears as well, leaving qaula C with a lengthened vowel- The State Of the string pairs, after composition of the alternation rules, is shown in Figure (9b Further composition af

•etelaxation" coles allowing the optional deletion of short vowels and the diacritics completes the picture. The final transducer will directly map from (_,_) or kibl ('Z.Æ) ar any partially voweled variation of the spelling to the upper side string [ktb&CaCaC]+Verb+FormI+Perf+Act+at+3P+Fem+Sg Jn the web demos 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; k*tabat

Upper: Ibny&CaCaCl ·Verb+FormI+Perf+Aci+at+3P4Fem•Sg
Lower: banat

Upper: lgwl&CaCuCl+Verb4Form14Perf+Act+a+3P4masc4Sg Lower:
qaala

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

### .5.1.1.3Advantages and Availability Of Finite-State Implementations

By keeping within the finite-state domain, grammatical components can be defined, combined andmodified using standard finite-state operations. Lexical transducers can be forwards to generate or backwards to analyze, and they are computationally very efficie]lt 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 codc that applies lexical transducers to input strings is also completely languageindependent. Thus the code that runs the Arabic morpho:ogical analyzer is exuctly the same code thet 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 heen licensed to over 70 universities nd non-commercial research centers.

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

3.5.1.2 Towards Understanding Arabic: Logical Approach for Semantics Fot 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 co Haddad and Yaseen (2003), these aspects are very important in the NLP. However, achieving Arabie understanding "requires actually a differentiated and deep semantic processing". Their project Towards Unders;runding Ar06ie: A Logical Approach For Semantre

Representation is directed to build a framework for processing the Arabic language in order to achieve the understanding of the language electronica] ly.

## 3.5.1.2,1 Arabic Understanding

*Artificial* Inre\digence has a long time ago recognized the necessity of performing some semantic rences.Haddad and Yascen 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 accordi".g to the researchers, might reside in the "'complexity of this field and in the invisible collaboration between Artificiål 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 panicularly 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, Semantic Processing

In semantic processing, different basic tasks havc to be performed at different levels. These tasks imply: semanttc composition (construction of semantic representation for capturing the semantic potential, of Aeabic propositions), semantic resolution (determining the current semantic value and the disambiguation onder using context knowledge) gnd semantic evaluation (extracting of relevant information based on performing 50me deductions and inferences on the semantic rcprcscntation 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 e natural language propositions in a specific domain. For example, concepts like $^{E'}$ (study), (project) and (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 tole for the whole semantic reasoning process, according to Haddad arid Yaseen.

### .5.1.2.3Semantic Representation

There are many reasons to choose a logical language as a rovger janguc.ge for the meaning representation. For Haddad & Yassen, logic represe]lts in particular a well-known rnæaning represen\ation formutism that differentiates between syntax and semantics, In addition, it enables inferences over quantified descriptions, which are basic requirements for act adequate meanino representation far any natural language

Fuahermore, representing Arabic sentences as logic programs have the unique of performing some semantic reasoning tasks on a code based on Arabic predicates. Therefore embodying logical formulas with Arabic predicates is a very inreres,ring 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)(∃x)( طالب (x) ∧(∃y)( يدرس (x,y)) )

(2.2) taleb(x) A yadrus(x, y)) )) (∃x)( student(x)

A (study(x, y)))

In there are important methodical principles and constraints for any semantic formalism designed the practical applications. These include Compoejness-, Modularity.

Generality. Expressive Power, Efficiency, Implementation Independence, Theory Independence,

Since Arabic is based on verb-noun and *noun-noun* opposition, we can establish a conespondeme between Arabic sentences and predicate logic first order (PLI) formulas. The verb or the " of an Arabic sentence (the nominal predicate of a nominal sentence) can be assigned to a predicate argument-structure of the corresponding PLI formula. The noun phrases ear. 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 Yascen argue that 'Sa 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 fdrrnal representation formalism, it does not provide any compositional methods. Based on the type rheory 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 senrenccs and are in general, inappropriate foc treating semantic processing oftexts 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 I-Calculus leads to z compositional framework that is able to capture such problems.

Our current view for achieving natural language understanding in the COäitext of the semantic representation of Arabic, according to Haddad and Yaseen (ibid), is to employ I-Absjraclion far constructing logicu/ formeelas acting as meaning representation for Arabic sentences. I-DRY semanlic conso•ucrion will he the next goal for achieving Arabic text understanding.

### 3.5.1.2.5 The Logical Form

Since Arabic d[5tinguishes betwæn different types Of sentences: Verbal Sengences (VS), *Nominal*Sentences (NS} and Copulotive Semences CCS) application of *λ-Calculus* requires a contextual interpretaiion of the meaning ofthe determiners in the different types of sentences. Because of the fact that the nominal and copulative sentences start with noun phrases, their seman}ie construction and representation woeld generally be similar to those sentences in Ehglish 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 Yaseet'i argue. The most used Arabic determiner “ " can be understood as a quantifier. Based on the standard analysis for determiners in the type theory, we can interpret the determiner det (l num: sing} as

{3.1) XP XQ$\exists\chi(\forall\gamma(P(\gamma\Leftrightarrow\chi=\gamma)\wedge Q(x))2$

where J Il denotes the meaning orthe determiner " .

The indications indefinite articles can be Interprewd as 2-quantifieß as follows

(3.2) IlIndefinite indication $\|\Rightarrow$: $\lambda P\ \lambda Q\ \exists\chi(P(x)\wedge Q(x))$

"J*' expresses that there exists only one thing of being P and Q, which impiies that the cordinalify Of p has to be I.

In general, 2 quantifier differentiates between two things: a restriction ami a scope (S), P(x) represents in 3, i the restricted sent *'Restriction* or Ike Base) and Q(k) the scope (the propos:iiion about the .resfricjed*sel*).

Generally a determiner can be exptessed

(3.3} as Il Detll IR IS (Quantifier(RS))

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

For example, the quantifier (most) expresses that the IRASI holds relatively a large portion of IR'.

By treating the " Icy-quantifier" and numeric quantifiers, we have adopted a similar concept presented in (Bir,ot, 1991, cited in the research paper) by introdueing the new quantifiers ID) expressing singular definite determiner, (IJ expressing plural definite determiner and (Jin: ) representing numeric defin ite determiner.

Determiners like and i" can be interpreted as all-qualifiers as followe

(3.4) II كل || ⇒ : λP λQ V(x, Q)

3.4 can also bc expressed based on 3.3 as follows;

(33) kPkQ( (R. S)) i" can be interpreted as

3-quantifier:

(3.6) || بعض || ⇒ : λR λS ∃x(x, RAS)

Adverbs (ظروف)modify verbs and adjectives and therefore they arc intentional like

quanti fiergt

(3.7) Il Adverb Il s: XP λQ(ظروف(P,Q)

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

(3.8) | Noun Il⇒ : λx

A noun means in 3.8 that there is something, which can have the property of being (noun), For example applying the meaning of the noun (student), Il 10 the proper name 04 (Ayman) means (hat there is somebody whose name if "*Ayman" with the property of being a

student:

|| طالب || ( أيمن ) ⇒ : λx طالب (x) ( أيمن )
⇒ : طالب ( أيمن )
⇒ : student (Avman) student (Ayman)

Adjectives can bc represented similarly:

(3.10) llAdjectivell⇒ : λ

## Verbs

Verbs in Arabic can be intransitive or transitive We can represent their meaning as follows"
(3.11) Illntrunsitive Verb] |⇒: λ x فعل1 (x)
(3.12) llTransitive Verbll ⇒: λ x λ.y فعل2 (x,y)
(3.13) ]Di.transitive Verbll⇒: λ x λ yλz فعل3 (x,y,z)

## 3.5.1.2*6 Compositional Rules

Ir'i 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 af NS can be obtained by applying the meaning of the to rhe meaning خبر

||H||.That means applying of
||M|| (||H||)

So if the " (M) consists or a determiner and a noun, as it is the case in the following incomplete Logic Grammar, then means the application of the meaning of the noun to IIDetll. The meaning of' the entire nominal sentence can then bc achieved by determining the

meaning "1 li خبر ||and its application to 11Mll,

<NS> → <M×H> sem ||NS||
<M> → <Det×N> - IIMII j
sem 11MlHlDetll(llnounll)

<Det→ ال / كل / بعض / …
Sem || ال || = λRλS( (x,R^S))|
|| كل || = λRλS( (x,R→S)|
|| بعض || = ….

(3, 14)<H>→<Noun>| <Adj>|.....
Sem IINoun Il —λx Noun
Il Adjil = ax Adj(x)

For example the meaning of the 14411 M (Det (J ,sfng), noun ( ) is the application oi the meaning of the noun to the meaning ofthe determiner

|| الطقس || ⇒: λRλS( (x,R^S))( || طقس ||)
(3.15) ⇒: λS( ال (x, طقس (x) ^S))

Applying the meaning of the adjective (nice), which takes the role or yields the meaning Of the sentence 'الطقس جميل 'i'

|| الطقس جميل || ⇒: λS (1ال (x, طقس (x) ^ S )) ( ||
λS( (x,R^S))( || ||)
(3, 16) ⇒: 1ال( (x, طقس (x) ^ جميل (x)) Il

Considering determiners as quantifiers requires the application of their meanings to the meaning of ather syntactical categosies, Since verbal sentences start with verbs, and if the
(subject) contains a determiner, the meaning oi the subject can be achieved by applying the meaning of the 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'

<VS>→<VerbXSubXObj>
Sem ||VS || = ||Sub|| (
(3.17)
(llObj311(llVerbll) )

Il VS] i.e. the meaning OF VS, is the application or llVerblto the meaning of the Object and eventually to l!Subl.

For example the meaning Of Jl (the student) the VS "يدرس الطالب الحاسوب"is
(3.18)ل طالب

λS(IJ (x, (x) S)) (110bjll (1B Verbll) )

Applying of the liObjll to the meaning of the verb yields;
(3.19) || يدرس الحاسوب || ⇒ :
1ل (y, حاسوب (y) ^ يدرس (x, y) )

Reyarding of(3.20) as the meaning of the scope of the determiner in (3.10) yields;

(3.20)1ل (x, طالب (x) ^ ل (y, حاسوب (y) ^ يدرس (x, y)))

(3.21) IJ (x, student(s) Idl {yt computer(y) 'h study (x,y)))

This re-search paper demonstrates that: First rept•sentiog Arabic sentences as logic programs has the unique facility of performing some serrantic 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 A-calculus for constructing logical formulas acting as meaning representation for Arabic sentences Third, extending this approach to ADRT leads to a good strategy for solving problQm5 involved in text anaphoric and a

modu lar composition, according to HGddad and Yaseem

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 section, the contribution of Arab universities and research institutes with regard to research and development of issues telated 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 bc 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;

l) 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 Ambicization, 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 fonvard was achieved when Said Haydet', a professor at the Montreal University (otiginalty 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 ، ك ، ل ).

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

- The Arabic language was used in the database and information retrieval systems.
- Software systems ',vere developed in Arabic. such as Basic and Logos,
- Preiim inary systems for the computational generation of Arabic language were developed.
- The develppment of partial systems for marphoEogical analysis,

- After ten years of discussions, the Unified Arab Code for electronic data exchange on the Internet received Arab unanimous agreement,
- 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 syntaetie analysis, diacritizati0fl and segmentation,
- The automation of Arabic dictionaries.
- The development of text analysis software which was used the morphological analysis of the Qura•an.

- ne development of spell-checking systemsr 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 motphologic.al pattems 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 multinode syntactic analyzers. Sakhr could develop extensive word I ISts and a body of 20,000 rules for Arabic grammar and syntax.

In the last ten years, the internationalization of the www arvd the proliferation of eornmunieztion tools in Arabic, as shown earlier, demonstrate the need for a large number or Arabic NLP applications. As result, more research activities have been Igunched to address more general areas of Arabic language processing, including syntactic analysis. machine iranslation. document indexing, information retrievaly etc, Research in Arabic speech processing has made significant progress due to "morc 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 futther progress [r. 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 tald the researcher that machine translation in the Arab world is still a field 00 be revealed even for most Arab intellectual elites. In the Arab world, debate over machine translation still cmcentrates 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 eommerciBJ sector of the advantages ih using machine translation in the Information Agee

According to Labib, there is to consistent and/or systematic machine translation research in the Arab countries. There are individual programs even within the borders of one Arab countrys and. it is even hard to scan such prograrns„ The Pan-Arab Translation center does not have a record for any machine translation programs or applications evailable in the Arab countries, according to Labib. The Pan-Arab Centre is preparing plan for a mechine translation program, but it is still in its preliminary stages.

As part of the present study, the researcher ttied to contact various research institutes in a number Arab countries, which are specialized in Information Technology Research and a couple of Arabic language academies to see 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 }inguistic 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 aeademy.

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 resexrch projects they are developing in context with machine translation or language technolocv and engineering. It was tzalized that almost all the companies in DIC are basically working in sales and marketing, whereas development and programming are taking place in Olhe: countries like Egypt and Jordan. In Egypt, Sakhr was very cooperative. Their research Centre, headed by Chalabi Ashraf, was wiling 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 interest in the topic; some of the intellectual elite still believe that the machine is "stupid" and cannot translate and oth«s do not realize the feasibil ity 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 hete in Sharjah in coordination with the Higher Colleges of Technology. However, these projects are confined to the domains or human translation. AUS is encouraging such activity. Prof. Raddawi is heading a project to be launched in FOLI (2004) on machine translation and interpretation.

The researcher also contacted the Dubai e-Governtnent 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 technoiogy and machine translation in some Arab countries in a book entitled

*ıslation*In The World: Towards Establishing 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

TheInstitute is executing a program for specia!ized machine translation in coordination with the European Union to translate medical texts. The program is called %KRAMED', It foilows the transfer technique. It is part of the European CATz program. The Institute of Electronic Research in CBiro is developing the Arabic part of the project.

The Institute is aiso building multi-lingual dictionary based 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 engi neering, 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,
- The development of a database for Standard Arabic Voice Recognition.

## 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 fot machine translation for Arabic "d French, But, fot onknown reasons thi5 program was not published.

## Syrian Scientific Research Centre for Information Technology

Food Khouri, a member of the Syrian Scientific Research Centre told the researcher via fax that in Syria, machi,ne translation research is still in its very preliminary stages, There a plan to establish centre for translation and [anguage processing affi)iaced with Damascus University, according to Khouri.

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

- A software for Arabic letter recognition. Two systems were developed:

1) A system which works on the 'VAX — l" and QIBM-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 Arabieizaton 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 deveioped a Machine Translation System 'Turjuman' which will be launched soon.

It is clear that Il'esearch and development activities in the field of Arabic LanguBge Technology und the Automatic Processing of Arabic language applications are still 'very few a.ncl sirnpie. Arab un iversities 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 wortd'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 appücations 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 termino!ogy among Arab countries, according to Raddawi (2002). "MT systems and software and contribute in thc process or standardization of Arabic technical terminology. Consistency can be reached through MT software if put on line and used by everyone" (Raddøwi,, forthcoming).

Examples'

User name

Password

Outbox S..-A.Y البريد

Folder

Toolbar

The Arab academia in general has not realized this need, except for individual efforts, But same 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 Transkation

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 pert ot- Al-Alamiah Group) has deveoped schernes for machine translation from Arabic-English and English-Arabic.

Over the last 20 yeztss Sakhr has reeiized the importance of Arabic Natural Processing as a starting point for Language Technology application. In that foundational approach, Sakht 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 IZligious instruction (Arabic versions of the

Holy Qura$^{i}$an, Hadith databases and Arabic tutorials), Internet front-ends, optical character recognition for scanning Arabic text-to-speech application5 and machine translation,

For NLP, Sakhr has developed software tools for morphological and syntactic analysis, diacritization and segment8(ion, 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-Eulish.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 successfulYy applied to handle English. Machine translation engine performance has beer boosted by 2 statistical language model contributing in the lexical and morphological disambiguation of the source language, in addition to enhanced word yelection on the target language, according to Chalabi, The Sakhr language statistical model is supported by two balanced corpuses one Fer English and another for Arabic 200 million words,

hi 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 persona] computers in the Arab world. la order to open the door to the Arabic user to efficiently Itse the [nternet, Sakhr has developed a number of tools and products will be listed iater.

### 3.6.3.2 Coltee

Established in 1990, Coltee 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 Ignguage 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 ami grammar checker used for the first time in 1 997 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 Markov Models (HMM)' (to extract the Arabic linguistic features required for Information Technology applications, according to Al-Sabah (2003),

Coltcc 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 ntme 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 ha5 appeared in multilingual mode and has translation memory.

### 3,6.3.4 Cimos

Cimas 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 traneslation and Arabicizatiorn services to be installed by other interested companies. Cimos has developed a number of machine translation software systems such as An-Nakei Al-Arabi. Al-Kan .4i.Mu 'wormetc.

### 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 wel)-knowj'i commercial narne "Al-War. Its first software was 'ÅLMuturjim Al' Arabey for professional translation. The company has recently developed translation *engine' which uses Artificial Intelligence solve the linguistic problems in translation, according to Ai Marzouki (20CQJ, the Director of Al-Marzouk For Techn0104C and Information, thc representative of ATA in Riyadh.

## 3.7 List of Commercial Arabic Machine Translation Software systems

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

Al-MutarjimAt•Aroöey; English — Arabic

Golden Al-Wafi v2.Off English —Arabic

Wafi V4·00: English Arabic

'utarif:English —Arabic — English

An-Nakel: English — Arabic — English

n-Nakel: One svey Arabic — English

An-Nekel: One way English —Arabie

An-NGkel Multilingual Translation sysienr,

## 3.8 Arabic Translation 'engines' on the Web:

The widely used Arabic Translation Web Portals

are; Tarjim, the Arabicization tool on Ajeeb,,com,

l-Mishar developed by ATA

CAT Translator: Bi—directional English — Arabic — English - Sakhr.

On-line Translation - Sakht.

Since companies like Sakht', 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), Thc 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 over five decades ago, A Ithough research in machine translation lacks an appropriate, consistent and easy to use criterion for evaluating the resuits. evaluation tools are indispensable in that they allow us to compare two translation systems or to information as to how a variation of any system might affect the quality of translation. Evaluation of MT system is required, both by developerss before and after 5YStern modifieaG0hS, and by end-users who wish 10 compare different systems before making a pulthase,

The quality of MT translation systems has currently being measured by usi ng a variety o' techniques and generally depends upon the context in which the MT system is being used. Whereas many other parameoers are relevant to the quality of the system. it is often the ourpt'f qua/i1Y 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 propo$ed 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.

4.4 Data evaluation

4.5Corpora analysis.

4.6 Conclusion

4.7 Recommendations

## 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 Hatirn and Mason (1990), one obvious application of linguigties 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 pmfessiom "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 syntectie structures had to be exchanged for TL structures; lexical items from each language had to be matched and the nearest equivalents selected.

According to Hatin-j & Mason (1990, p.22)

> While 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 ands 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" (Hatiitl & Mason, 1 990! p.2S), Morphology and syntax constituted the main areas of analysis.

Since meaning is the heart of the translator's work, it follows thal 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 eontext of situation in which language activity takes place, However, such attempts are very recent in MT and have not achieved moch, In Arabic computation, debate is still ongoing regarding syntax and morpha10Ü,
This point of interest will be investigated in the corpora analysis.
c) Chomsky's generative — transformational model analyzes sentences into a series or related levels governed by rules. The key features of this model can be summarized as follows:

I) 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.

TIR structure relations described in this model are held by Chomsky to be a universal feature of human language. Chomsky's mode]. as was mentioned before, is the basis upon whichcomputational linguistics built,

d) Nida and Taber (1 969, p, 39) claim Illat 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 transfet, 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.

t) Again, according to Nide, 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 ort 'four basic requirements of a translation:

l) Making sense;

2) Conveying the spirit and manner of the original.

3) Having a natoral 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 thc linguistic servitude's of the TL" (Vinay and Darbelnet, 1958 p.S6).

h) According to Vinay and Darbelnet, unacceptable message in translation when translated literally means.' l- gives another meaning, or 2- gives no meaning, or

3- strucLurally impossible, or

4- does not have a corresponding expression within the metalinguistic experience ofthe

5-has a corresponding exptession, but not within the same register.

## 4.2. Points or Reference in MT

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

- MT is by different users for different nccds- Users' needs usually determine MT output; whether the user needs a good-polished translation just gist translation,
- MT is successful in technical and scientific texts, It is good in transiatjng specifie domain area. It is not successful in literary texts.
- MT uses various transiating strategies, among which transfer strategy is mostly used in Arabic software programs. Transfer is a three stage strategy where: Ij 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

Th is section prepares the scene for the corpora analysis.

### 4.3.1 Language Combination

In that the field of MT systems awalysis is 50 broadv the scope of this study will focus on the single language pair; English - Arabic,

### 4.3.2 Text Types

Four different samples bave 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 (In formation 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 overall idea about the main operational and functional ideas in the text. They need to have accurgte information about "prototype mts from the web metrics test bed" etc. The users expect accuracy and have e•nugh clear information. Users of the political text (a news article from thc 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 jnay be enough for most political readers,

### 4.3.4 MT Systems

Two commercial systems were randomly selected from the market: 'Al-Wafi•version 4[7] developed by ATA company based in London, and "Al-KafV developed by Cimos Company

Based in Paris. Both systems are developed to be used byr 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 eval uation procedure is based on MT users need5* 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$^{i}$ needs. Two types of "iteria have been selected, one at the sentence level, the other at the text level*

l) Analysis wil! begin at the sentence level by checking: syntax, morphology lexis.

2) The overall text will then be evaluated to cheek 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-Rafi. The analysis will begin with two medical texts, followed by a technical text and finally political text.

### 4.541 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 compeny in order to advertise this new produet,, The readership of such texts can he doctors, pharmacists or patients. fn all cases, the trensiation is expected to be clear and accurate it is essential.ly user oriented.

BRAND NAME: Symmetrei

DRUG CLASS AND MECHANISM: Amantadine is a synthetic (man-made) anti-viral drug 'that can inhibit the replicatiM1 of viruses in ceils, To prevent a viral infection, the drug should be present beface exposvre to the virus, Clearly, this is not practical for most viral infections. It was initially used 10 prevent A during flu season, and, if given wihin 24 to 48 hours of the onset of symptoms, to decrease the severity of the flu, Later amantadine was found to cause improvement in the symptoms of Parkinson$^{s}$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, tteurotranmitter in the braiß, that is reduced in

Parkinson's disease. Arnantadine is less effective than levodopa in Parkinson's disease but can offer additional benefit when taken with [evodopa- Amantadine is less effective than levodopa in Parkinson's disease "t can offer additional benefit when taken with levodopa, Amantadine was approved by the FDA in 1966GENERIC 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 5 and 30$^{2}$C (59 and 86$^{0}$F).

PRESCRIBED FOR: Amantadine is used for the prevention or treatment of infections with inflceoza A virus, especially for individuais at high-risk such as immunosuppressed patients and nursing home residents, It should not be used asa substitute for vaccinatio, Amantadine also is used for control of the symptoms of Parkinson's disease,

DOSING; Amantadir,e is taken once or twice daily with or without food, If it causes an upset stomach, it can be taken with food,

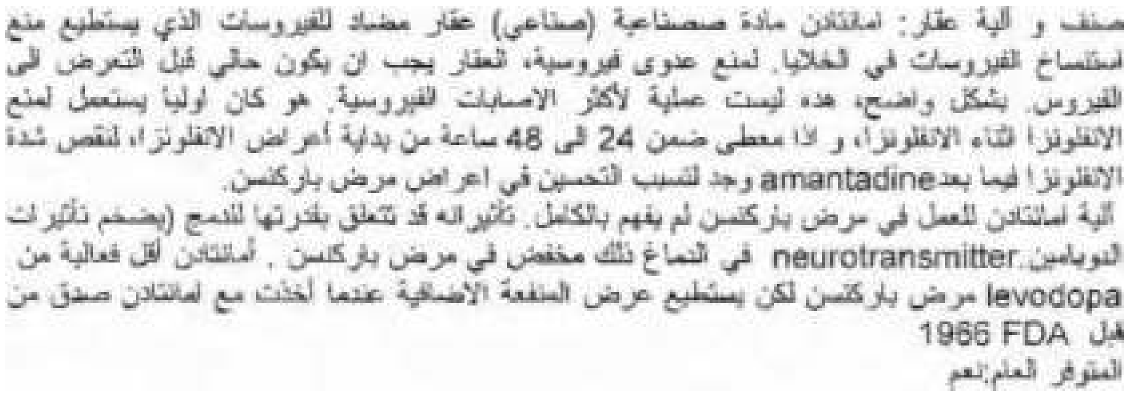
"For treatment of innuenza, amantadine should be strated within 24 to 48 hours after the onset of signs or symptoms, TO puevent influenza, amantadine should be started as soon as possible after exposure to the influenza virus and continued for at least 10 days.

Persons with reduced kidney function and elderly persons may need tower doses (or less frequent doses),

Al-Wa fi

The translated version

العلامة التجارية سيميتريل



صنف و آلية عقار: امانتادن مادة صصناعية (صناعي) عقار مضاد للفيروسات الذي يستطيع منع استنساخ الفيروسات في الخلايا. لمنع عدوى فيروسية، العقار يجب ان يكون حالي قبل التعرض الى الفيروس. بشكل واضح، هذه ليست عملية لأكثر الاصابات الفيروسية. هو كان اوليا يستعمل لمنع الانفلونزا اثناء الانفلونزا، و اذا معطى ضمن 24 الى 48 ساعة من بداية أعراض الانفلونزا، لنقص شدة الانفلونزا فيما بعدamantadine وجد لتسبب التحسين في اعراض مرض باركنسن.

آلية امانتادن للعمل في مرض باركنسن لم يفهم بالكامل. تأثيراته قد تتعلق بقدرتها للدمج (يضخم تأثيرات الدوبامين.neurotransmitter في الدماغ ذلك مخفض في مرض باركنسن . أمانتادن أقل فعالية من levodopa مرض باركنسن لكن يستطيع عرض المنفعة الاضافية عندما أخذت مع امانتادن صدق من قبل FDA 1966

المتوفر العام:نعم

التحضيرات: امنتادن متوفر ك 100 gelatin mg كبسولات ناعمة و بينما عصير يحتوي MG50 لكل ملعقة شاي

الخزن: مخزن في درجة حرارة الغرفة بين 10 و 30 C (59 و 86 F)

وصف ل: امنتادن مستعمل للمنع أو معالجة الاصابات بالانفلونزا الفيروس خصوصا للأشخاص في خطر جدا مثل immunosuppressed سكان مركز رعاية المسنين و المرضى و هو لايجب أن يكون مستعمل كبديل للتطعيم. امنتادن أيضا مستعمل لسيطرة أعراض مرض باركنسن

المداواة: امنتادن مأخوذ مرة أو مرتين يوميا مع أو بدون الغذاء إذ يسبب معدة مضطربة، هو يمكن أن يأخذ بالغذاء

لمعالجة الانفلونزا،amantadineيجب أن يبدأ ضمن 24 الى بعد 48 ساعة بداية الاشارات أو الاعراض و يجب أن تستمر ل 24 الى بعد 48 اختفاء الاشارات أو الاعراض لمنع الانفلونزا amantadine يجب أن يبدأ بأسرع ما يمكن بعد التعرض الى فيروس الانفلونزا و المستمر الى 10ايام على الاقل.

اشخاص بوظيفة الكلية المخفضة و الاشخاص المسنين قد يحتاجان جرع اوطى (أو جرع اقل تكرارا)
تفاعلات عقار: يضيف امنتادن الى كحول تأثيرات التسكين و مخدرات التسكين الاخرى مثل الصنف benzodiazepineقلق (و مثال على ذلك: - فاليوم، اتيفان، كلونوبين، كسانكس، امبين) ، صنف tricyclic لمضادات الكآبة ( مثال على ذلك: الافيل، توفرانيل، نوربرامن) dicyclomine(بنتيل)، بعض المضادات للهستامين (بنادريل، فيستاريل، تاراكس، تافيست) يخدر agonists (ومثال على ذلك: دلودد، فيكودن، بيركوسيت، كودين) و بعض الادوية ضد ارتفاع ضغط الدم (و مثال على ذلك: كتابريس، اتدرال) مثل هذه المجموعات تستطيع تسبب الدوخة، تشويش، lightheadedness ، اغماء ، او دوخة على الموقف.

## 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 sticks to the source text, but in accordance with the rules and confinements of the target language, In word-for-word translation, the target text goes with the source text following all its rules and structures, Henczs the translation becomes odd and the message is usually lost. What makes things worse in the two translations of this partieolar medical text is that translation is done between two incongruent languages, i.e.

English and Arabic.

l) As a result of thc word-for-wold strate•o 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 ...

- Il was initially used to prevent.....

- هو كان أوليا يستعمل لمنع الانفلونزا

2) Parsing: it is difficult to conduct parsing in such a confüsed 5entence structure, It is impossible to consider cohesion here. However, and for the sake of checking the work ot 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 حالي.. , instead of يكون حاليا Inother

similar case, the parsing was eorreetly done:

This indicates the luck of rules and consistency in passing.

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...

امانتادين مادة صناعية ... عقار مضاد الذي يستطيع منع ...

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

Examples,

a) There is clear inconsistency in pronoun substitutions tin 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 المسنين قد يحتاجان instead of الأشخاص المسنون قد يحتاجون...

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

- تسبب الدوخة، تشويش، إغماء..

It should be either

- تسبب الدوخة، التشويش والإغماء ..

- تسبب دوخة وتشويش وإغماء.

[t is noticed that Al-Wafi system succeeded some cases offsyntax and parsing and failed in others. This is somehow strange since it is assumed [hat the parser and the syntactic analyzer in uny machine translation system are built according to the rules of the TL so that the systern respects such rules and the trans;ation reads natural, Al-Wai' 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, tot instance, Example sentences can be included in the built up of the analyzer for better application in the course of tran51 ation.

## Morphotogy

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

1• Passive voice confirms the occurrence of inaccuracy in the translation output of this text,

Exatnples of passive verbs wrongly formed:

- *' if givef is wrongly transferred into Arabic instead of 131 "

- The verb "is used for" is transferred into instead of" يستعمل"

- The imperative verb form store at room temperature is again transferred into a passive " مخزن"instead of" يخزن "

The passive voice in the above mentioned examples (case 3), should be extracted from the tri-root( فعل + ي ) or ( ) to have 'LA as é*iand as 0>4.

2- There is also a ptoblem in plural formation: In the sentence

- الأشخاص المسنين .. جرع أوطأ ...

The word here is meant to be a broken plural of جرعة.But according to Arabic rules جرعshould take the natura\ feminine plutålجرعات

However, all other plural forms are correctly generated according to the ruless even the broken plural like سكان, أماكن, أعراض and others.

It is 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 lexica] items which usually have various retérential meanings, but whose usage differ according to context. •nits is where the role of the human interaction plays an important part in selecting 5uitable meaning$ It demonstrates the demand for understanding the pragmatic constituents where various technical items becomes clear here

Examples:

l) „,-drug that can inhibit the replication of viruses in cells. Replication is translated here as [illegible], The intended meaning here is Moreover, has now acquired a fixed meaning cloning.

2) ..the drug should be present before exposure to the virus, Present here is literally translated as What it is intended here is .

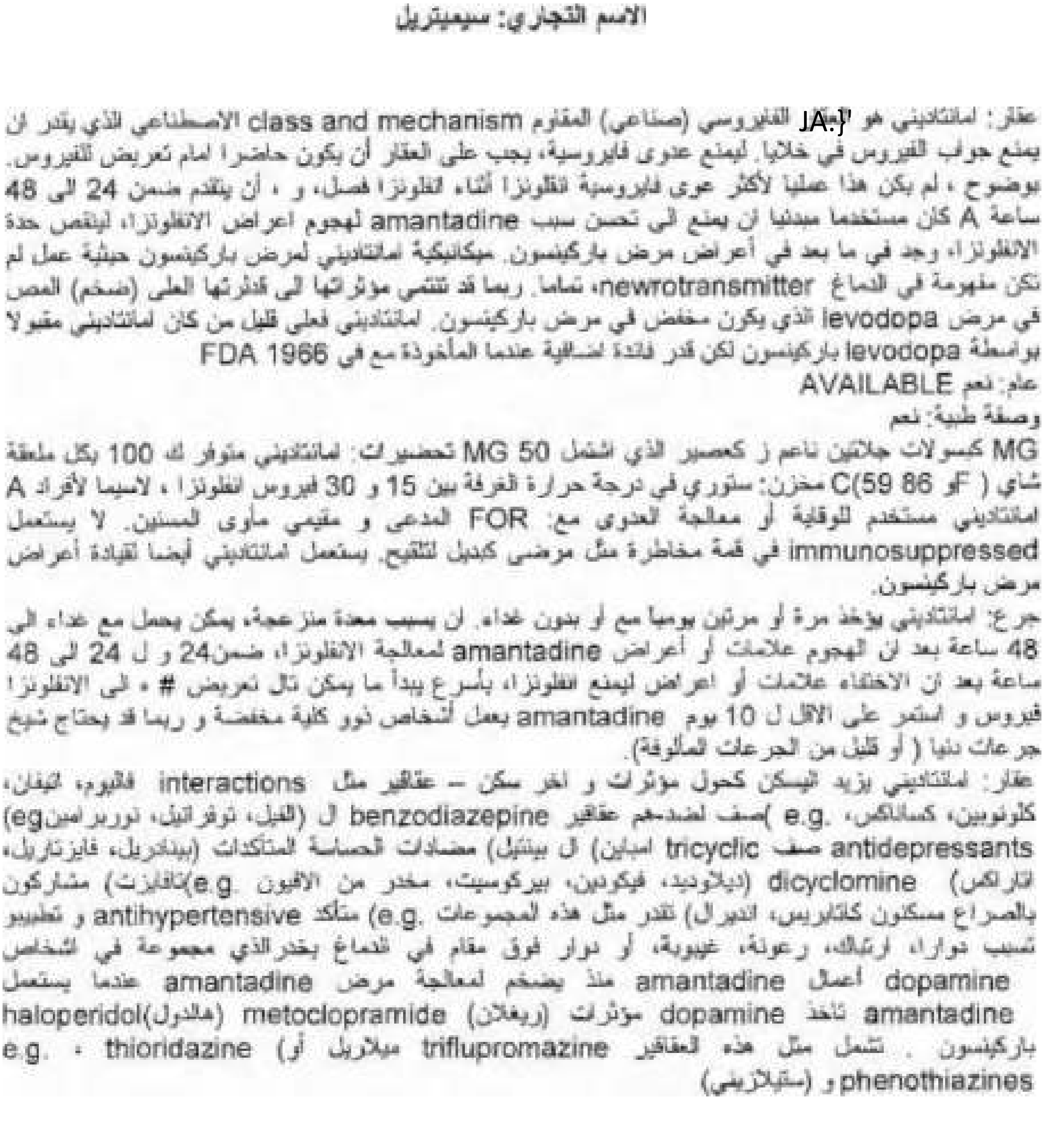
t fits

3) ...to decrease the severity of the flu. Decrease is tran51ated as Ä, while What is meant is _

In brief, the word-for-word translation of the previous is a good model of what Hutchins calls "unnatural literalness". The translation closely adheres to the source language structute 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



الاسم التجاري: سيميتريل

عقار: امانتادين هو JA.}' الفيروسي (صناعي) المقاوم class and mechanism الاصطناعي الذي يقدر ان
يمنع جوف الفيروس في خلايا. ليمنع عدوى فايروسية، يجب على العقار أن يكون حاضرا امام تعريض للفيروس.
بوضوح ، لم يكن هذا عمليا لأكثر عدوى فايروسية انفلونزا أثناء انفلونزا فصل، و ، أن يتقدم ضمن 24 الى 48
ساعة A كان مستخدما مبدئيا ان يمنع الى تحسن سبب amantadine لهجوم اعراض الانفلونزا، لينقص حدة
الانفلونزا، وجد في ما بعد في أعراض مرض باركينسون. ميكانيكية امانتادين لمرض باركينسون حيثية عمل لم
تكن مفهومة في الدماغ newrotransmitter، تماما. ربما قد تنتمي مؤثراتها الى قدرتها العلى (ضخم) المص
في مرض levodopa الذي يكون مخفض في مرض باركينسون. امانتادين فعلي قليل من كان امانتادين مقبولا
بواسطة levodopa باركينسون لكن قدر فائدة اضافية عندما المأخوذة مع في FDA 1966
عام: نعم AVAILABLE
وصفة طبية: نعم
MG كبسولات جلاتين ناعم ز كعصير الذي اشمل MG 50 تحضيرات: امانتادين متوفر ك 100 بكل ملعقة
شاي ( Fأو 86 59)C مخزن: ستوري في درجة حرارة الغرفة بين 15 و 30 فيروس انفلونزا ، لاسيما لأفراد A
امانتادين مستخدم للوقاية أو معالجة العدوى مع: FOR المدعى و مقيمي مأوى المسنين. لا يستعمل
immunosuppressed في قمة مخاطرة مثل مرضى كبديل للتلقيح. يستعمل امانتادين أيضا لقيادة أعراض
مرض باركينسون.
جرع: امانتادين يؤخذ مرة أو مرتين يوميا مع أو بدون غذاء. ان يسبب معدة منزعجة، يمكن يحمل مع غذاء الى
48 ساعة بعد ان الهجوم علامات أو أعراض amantadine لمعالجة الانفلونزا، ضمن24 و ل 24 الى 48
ساعة بعد ان الاختفاء علامات أو اعراض ليمنع انفلونزا، بأسرع يبدأ ما يمكن تال تعريض # ، الى الانفلونزا
فيروس و استمر على الاقل ل 10 يوم amantadine يعمل أشخاص ذوو كلية مخفضة و ربما قد يحتاج شيخ
جرعات دنيا ( أو قليل من الجرعات المألوفة).
عقار: امانتادين يزيد اليسكن كحول مؤثرات و اخر سكن – عقاقير مثل interactions فاليوم، اتيفان،
كلونوبين، كسالاكس، e.g. )صنف لضدخم عقاقير benzodiazepine ل (الفيل، ترفرانيل، نوربرامين eg)
antidepressants صنف tricyclic امباين) ال بينتيل) مضادات الحساسية المتأكدات (بيناجريل، فايزتاريل،
اتاراكس) dicyclomine (ديلاوديد، فيكودين، بيركوسيت، مخدر من الأفيون e.g.)كافليزت) مشاركون
بالصراع مسكنون كتاليريس، انديرال) تقدر مثل هذه المجموعات e.g.) متأكد antihypertensive و تطبيع
تسبب دوارا، ارتباك، رعونة، غيبوبة، أو دوار فوق مقام في الدماغ يخدر الذي مجموعة في اشخاص
dopamine أعمال amantadine منذ يضخم لمعالجة مرض amantadine عندما يستعمل
amantadine تأخذ dopamine مؤثرات (ريغلان) metoclopramide (هالدول)haloperidol
باركينسون . تشمل مثل هذه العقاقير triflupromazine ميلاريل أو) thioridazine ، e.g.
phenothiazines و (ستيلازيني)

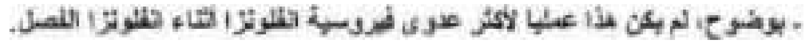
## Analysis

The translation output ot the Al-Kati 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 tight side text alignment. Seeond, the system is unable to read end 'understand' words, so it) many cases the are either written in English or transliterated using Arabic characters

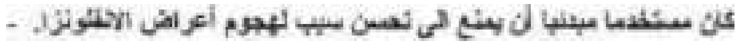
## Syntax

Cinguistically speaking, the whole sentence structure is confused, No system is followed in translating. [t is not even a word.for-word translation. There is no word order and the structure of the sentences are either mixed up, or the sentence5 are incomplete.

Examples ofthe jack of structure in word order;



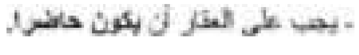
- بوضوح، لم يكن هذا عمليا لأكثر عدوى فيروسية الفلونزا اثناء الفلونزا الفصل.



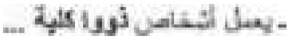
كان مستخدما مبدئيا أن يمنع الى تحسن سبب لهجوم أعراض الانفلونزا. -

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 persing are generally correccly conducted: example

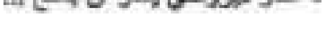


- يجب على العقار أن يكون حاضرا.

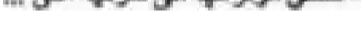


- يمثل أشخاص ذووا كلية ...

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



- عقار فيروسي يقدر أن يمنع ...



- تنتمي مؤثراتها الى قدرتها على ...

## Morphology

The system proved successful in the formation of inflectional and derivational words from their roots 01' sterns.

Examples;

1- (فيروس، فيروسية، فيروسات، فيروسي all derivations of j. Though the word is not originally Arabic, it has acquired the derivational rule of Arabic words2- From the tri-root the following words were derivedعالج، يعالج، معالجة، علاج

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

## Lexicon

The system poses serious lack of •understanding' of the word meaning.
I- fn some cases even the referential meaning is lost. Example; the replication of virus is transferred as(جواب الفيروس).

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

3- A phrase like elderly persons may need lower doses is transferredقد يحتاج شيخ جرعات دنيا.

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

In brier Al-Kars 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 contmue reading after the first two sentences

## 445.2 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, is published an the Common Cold Inc. site, The site's aim, as announced* is co 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 contaminete the hands of people 'With colds as result of nose blowing, covering sneezes, and touching the nose. Alsor 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 mother peräon during periods of brief contact. Also, cold viruses readily transfer 10 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 decidcs 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 neeesstry here to serve the objectives of the this part or the thesis.

## Al-Wafi

**Translated version**

### كيف البرد تنتشر

تنمو الفيروساتُ الباردةُ بشكل رئيسي في الأنفِ حيث يُضاعفون في الخلايا الأنفية وموجودة في الكمياتِ الكبيرة في السّائل الأنفي من الناس بالبرد.

التركيزُ الأعلى للفيروس البارد في الإفرازاتِ الأنفية تحدثُ أثناء الأيام الأولى الثلاثة منْ العدوى. هذا عندما أصابَ الأشخاصُ معديون جداً.

الفيروساتُ الباردةُ قدْ أحياناً تكُون موجودة في القطراتِ التي مطرُودة في السّعال والعطس.

الإفرازاتُ الأنفية التي تحتوي فيروسات باردة تلوّثُ الأيدي من الناس بالبرد كنتيجة للأنفِ يَنفخ، يغطّي العطس، ويمسّون الأنفَ. أيضاً، فيروسات باردة قدْ تلوّثُ الأجسامَ والسّطوح في بيئة مُعاني بارد. الأطفال الصّغارُ الخزانُ الرئيسي للفيروساتِ الباردة ومصدر جيد جداً منْ الفيروس الذي يَحتوي إفرازاتَ أنفية.

بيّنتْ التجاربُ بأنّ فيروساً بارداً يُحوّلُ بسهولة منْ الجلد وأيدي مُعاني بارد إلى الأيدي وأصابع الشخص الآخر أثناء فتراتِ الإتصال القصير. أيضاً، تُحوّلُ فيروساتُ باردة بسهولة إلى الأيدي كنتيجة للمس لوثتْ الأجسامَ والسّطوح.

الفيروسُ على الأصابع مُحوّلة في الأنفِ والعين بالإصبع لإلتقاط ولمس إلى الإتصال العيني. الفيروسُ أودعَ في العين يَهبط قناة الأنف فوراً في الأنفِ. مرة في الأنفِ، فيروس بارد منقول منْ قبلْ عمل mucociliary إلى المنطقةِ الغدية حيث تبدأ برداً. في بَعض الحالاتِ، فيروس بارد، الذي يُطرَدُ في الهواء في السّعال والعطس، قدْ يَهبط في الأنفِ أو يرى ويُسبّبُ عدوى.

## Analysis

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

l) ne misunderstanding of adjectives as verbs, Examples:

a) infected persons is ttansiated as أصاب الأشخاص instead of الأشخاص المصابين b)contaminated objects is translated as لوث instead of 3401 .

Whereas the cold sufferer is translated as باردة instead of الشخص المصاب بالبرد.

2) The system fol lowed the SL usage 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 المصابين بالبرد,

b) .,,to the adenoid area where it starts a eold is translated as الى المنطقة الغدية حيث تبدأ بردا instead Of الى المنطقة الغدية حيث تبدأ الاصابة بالبرد.

3) Illere is inconsistency among conjunctions, Examples

a) حيث يضاعفون .. وموجودة instead Of حيث يتضاعفون ويتواجدون.

b} يغطي العطس ويمسون الأنف insteadof يغطون العطس ويمسون الأنف.

4) The passive voice is used where the active voice is needed, Examples:

2) .where they multiply is translated as حيث يضاعفون instead of يتضاعفون

b} a cold virus readily transfers is translated as فيروسا باردا يحول بسهولة...instead off يتحول

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 البرد تنتشر

Such syntactic problems indicate che 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 example, 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,

## Morph010kY

The system succeeded co follow the morphological rules ofArabic in the formations of natural and broken piurals (such as الفيروسات، الافرازات، الأيدي، الأصابع ).the superlative form (such as الأعلى).the formation of verbs according to their patterns and their position in the sentences (تنمو، يتعاطون، يبلث، أصاب)(keeping in mind how the systems 'understands' the verb) and the formation of passive forms (such as مطرود، منقول، موجود).

## Lexicon

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

l) Cold viruses is as الفيروسات الباردة instead of البرد

2) , , -droplets that are expelled in the coughs is translated as القطرات المطرود inste2d of القطرات التي تطرح مع ...

3)The word eye is sometimes correcfly tyanslated as and other time it is translated as يرى.

4) ...children are the major reservoir of eold viruses is translated as الأطفال الصغار الخزان الرئيسي.The word is not a good selection in this context. A better selection J believe is المخزن الأساسي لتجمع ....

5) The expressions: finger-to-nose and finger-to. eye contact is wrongly transferred as بالاصبع لاشتمام ولمس الى الاتصال العيني And once in the nose is also wrongly transferred as ة في

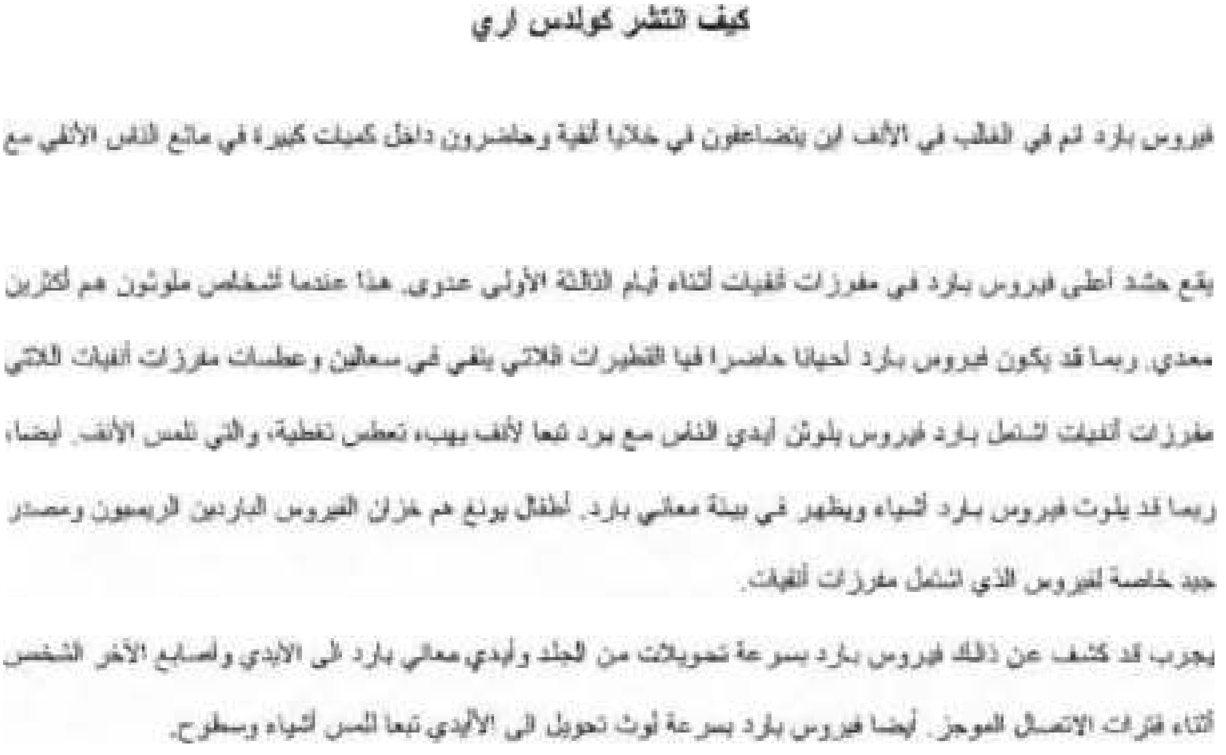
6) The system failed to 'recognize' the meaning of the medical term mucocitliary and kept it in English.

On of the problems of meanings I assume* is that the database include one derivational meaning, as it is the case with the word 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 reeds 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-Kati

The translated version

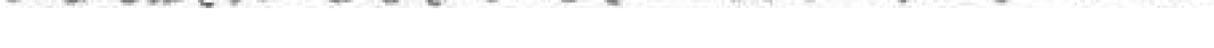


كيف انتشر كولدس اري

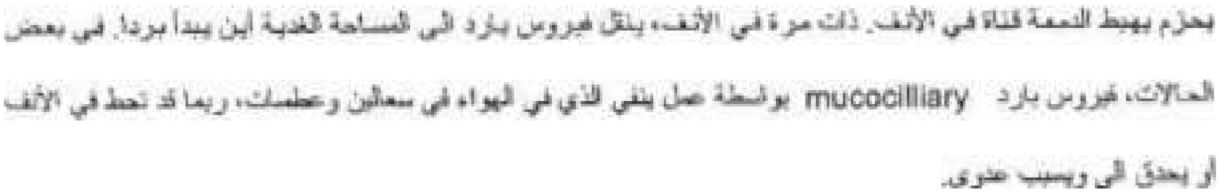
فيروس بارد ثم في القلب في الأنف اين يتضاعفون في خلايا أنفية وحاضرون داخل كميات كبيرة في مائع الانف الأنفي مع

يقع حشد أعلى فيروس بارد في مفرزات أنفيات أثناء أيام الثالثة الأولى عدوى. هذا عندما أشخاص ملوثون هم أكثرين
معدي. ربما قد يكون فيروس بارد أحيانا حاضرا فيا القطيرات اللاتي يلفي في سعالين وعطسات مفرزات أنفيات اللاتي
مفرزات أنفيات تشتمل بارد فيروس يلوث أيدي الناس مع برد تبعا لأنف يهب، تعطس تغطية، والتي تلمس الأنف. أيضا،
ربما قد يلوث فيروس بارد أشياء ويظهر في بيئة معاني بارد. أطفال يونغ هم خزان الفيروس الباردين الرئيسيون ومصدر
جيد خاصة لفيروس الذي اشتمل مفرزات أنفيات.

يجرب قد كشف عن ذلك فيروس بارد بسرعة تحويلات من الجلد وأيدي معاني بارد الى الأيدي وأصابع الآخر الشخص
أثناء فترات الإتصال الموجز. أيضا فيروس بارد بسرعة لوث تحويل الى الأيدي تبعا لمس أشياء وسطوح.



فيروس على الأصابع ينتقل في الأنف وعين بواسطة أصبع-على-أنف وأصبع-على-عين اتصال أودع فيروس العين داخل



بحزم يهبط الدمعة قناة في الأنف. ذات مرة في الأنف، ينقل فيروس بارد الى المساحة الخلفية أين يبدأ بردا. في بعض الحالات، فيروس بارد mucociliary بواسطة عمل ينفي الذي في الهواء في سعالين وعطسات، ربما قد تحط في الأنف أو يحدق الى ويسبب عدوى.

## Analysis

The translacion of the text demonstrates similar draWbacks as it was the case in the previous translated versions, with slight improvement in certain piaces where the structure of Arabic language is followed, However, it is still difficult to examine che 'IT structure since the word-fOt word policy was adopted in the translation of text without even respecting the TL linguistic roles. Syntax is one example,

## Syntax

It is impossible to follow the syntactic rules of Arabie here sinee the •whole structure ofthe text follows thc English structure. Parsing of course is impossible since there is not clear structure, There is no consistency for example, in parsing among the modifiers arid the modified (such as (البارِدين).The sysæm failed even to build simple structures such as الشخص الآخر (it is used as (الآخر الشخص) and القناة الدمعية is used as (الدمعة قناة ).

There is on [y one sentence which the system could build according to the Arabic structure (VSO): ... يقع خلف أعلى.
There are some cases where the system could achieve consistency in القطيرات اللاتي for example and مفرزات أنفيات (taking into consideration the consistency of gender regardless whether the derivations are correct or not)-

## Morphology

The system could achieve some success the formation of some morphological structures iike verb and noun formations, such as يتضاعفون، يلع، يتلقي، اشتمل، كميات، تحويلات، معاني، أيدي. However,

the formation ofsome verb and forms seems very odd.

Examples:

1 instead of

2) instead of JA" ات.

instead of and

4 instead Of .

## Lexicon

The worst part of the translation of this text is the translation of meaning. Aithough 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 ifthe meaning is there, the selection of the synonyms is not accurate.

Examples:

1) ..nasal fluid of people is translated as مائع الناس الأنفي instead of السائل الأنفي لدى ... 2) ...infected persons is translated as ملوثون instead of مصابون

3) cold virus is translated as بارد

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

5) Highest concentration is translated as insteadof عال

6) tn the last sentence, the word land in cold virus . ...may land in nose... is translated as The translation is correct, but the Arabic symrtym is too strong in this context,'åfl is a better translation, I believe.

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

S) The medical term rnucocilliary is kept as it is in English. This term is a compound of muco cilliary, This term is not recognized by both Al-Wafi and Al-Kan .

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

In general Al-Kafi has failed in most eases to respect the linguistic rules of Arabic language. What is astonishing in fact is its clear failute in the lexicon part for various reasons: first, the system is supposed to include a variety of dictionaries, among them the medicat lexicon. Second, the system supposedly includes databases that contain references for thousands of words. The words of this texts arc simple and common and hence, they should be part of any database.

The text needs a great effort form human translator or an editor to make it acceptable for dissemination. For information assimilation, t,he 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/lnformation Technology

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

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 he readable in order to adequately meet the users' need; the assimilation or information,

## Technical Overview

Coad usability is critical to the success of a website Usability evaluation has traditionally been a slow, labor-irtensive process which makes it diffcult 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 wchsite 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 8 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 rapidi remote, and automated testing and evaluation af website usability, The prototypes are used support the usability engineering research of the Visua)itation and Usability Group (VUG). A5 part of tbe Information Access Division of the In farmation Technoloo Laboratory at the National Institute of Standards and Techn0100,

VLJG :ncourages indusu•y 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 VisVtP,

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.

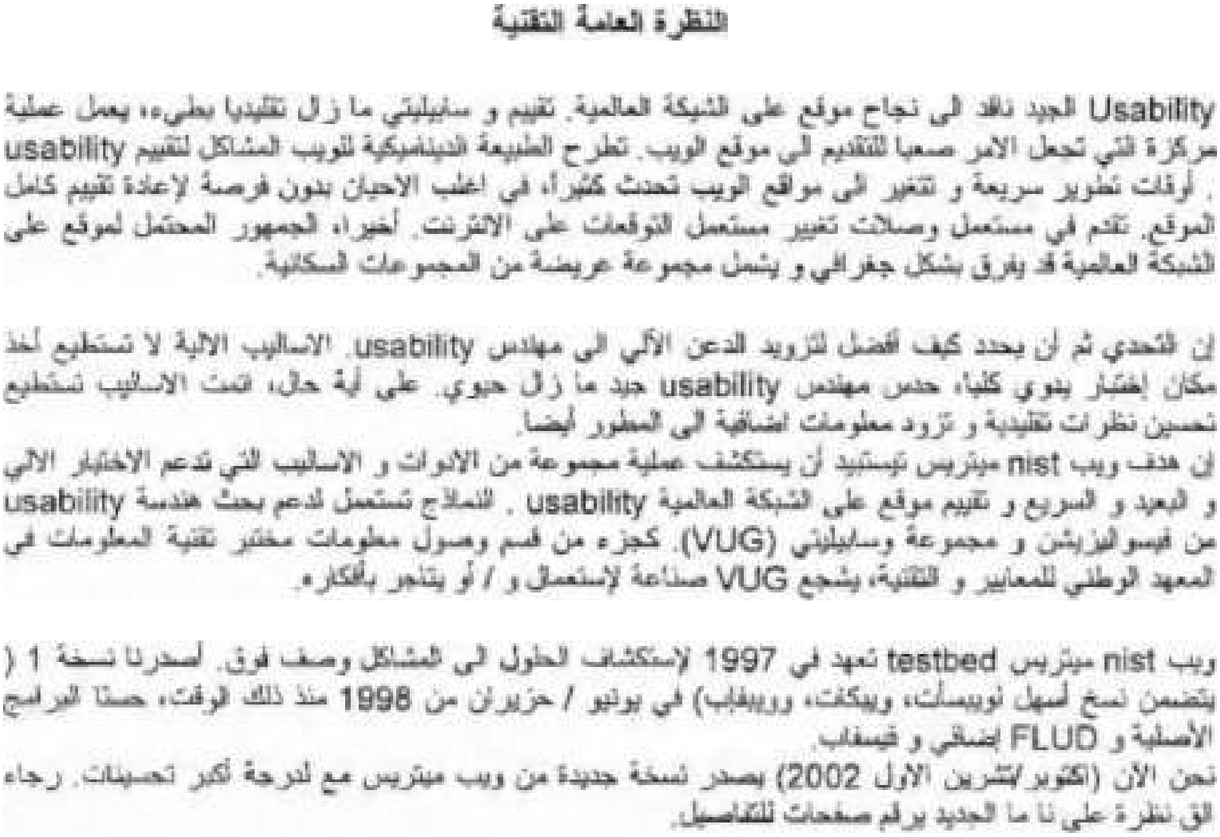
Prototype Tools from the Web Metrics Testbed — DOWNLOAD

TheWeb Static AnaEyzer Tool (WebSAT) checks the HTML of web pages against usability guidelines. either its ora 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 coreduct a simple category analysis across the It is a variation upon traditional card soaing techniques. The usability engineer establishes a set of categories and a number of items which areto 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



النظرة العامة التقنية

Usability الجيد نافذ الى نجاح موقع على الشبكة العالمية. تقييم و سابيليتي ما زال تقليديا بطيء، يعمل عملية مركزة التي تجعل الامر صعبا للتقديم الى موقع الويب. تطرح الطبيعة الديناميكية للويب المشاكل لتقييم usability . أوقات تطوير سريعة و تتغير الى مواقع الويب تحدث كثيرا، في اغلب الاحيان بدون فرصة لإعادة تقييم كامل الموقع. تقدم في مستعمل وصلات تغيير مستعمل التوقعات على الانترنت. أخيراء الجمهور المحتمل لموقع على الشبكة العالمية قد يفرق بشكل جغرافي و يشمل مجموعة عريضة من المجموعات السكانية.

إن التحدي ثم أن يحدد كيف أفضل لتزويد الدعن الآلي الى مهندس usability. الاساليب الالية لا تستطيع أخذ مكان إختبار يدوي كليا، حدس مهندس usability جيد ما زال حيوي. على أية حال، تمت الاساليب تستطيع تحسين تظرات تقليدية و تزود معلومات اضافية الى المطور أيضا.
إن هدف ويب nist ميتريس تيستيد أن يستكشف عملية مجموعة من الادوات و الاساليب التي تدعم الاختبار الالي و البعيد و السريع و تقييم موقع على الشبكة العالمية usability . النماذج تستعمل لدعم بحث هندسة usability من فيسواليزيشن و مجموعة وسابيليتي (VUG). كجزء من قسم وصول معلومات مختبر تقنية المعلومات في المعهد الوطني للمعايير و التقنية، يشجع VUG صناعة لإستعمال و / أو يتاجر بأفكاره.

ويب nist ميتريس testbed تعهد في 1997 لإستكشاف الحلول الى المشاكل وصف فوق. أصدرنا نسخة 1 ( يتضمن نسخ أسهل نوبيسأت، ويبكت، وويبفاب) في يونيو / حزيران من 1998 منذ ذلك الوقت، حسنا البرامج الأصلية و FLUD إضافي و فيسفاب.
نحن الآن (أكتوبر/تشرين الاول 2002) يصدر نسخة جديدة من ويب ميتريس مع لدرجة أكبر تحسينات. رجاء الق نظرة على نا ما الجديد برقم صفحات للتفاصيل.

ادوات نموذج من الويب ميتريس تيستفيد – انزال اداة محلل الويب الساكنة (ويبسات) يفحص إتش تي ام ال صفحات الويب ضد تعليمات usabulity، لما له، أو مجموعة IEEE سنة 1999-2001 تعليمات. هو يستطيع فحص صفحات فردية أو كامل الموقع على الشبكة العلمية.

اداة تحليل صنف الويب (ويبكات) يترك مهندس usability يبني بسرعة و يجري تحليل صنف بسيط عبر الويب. هو إختلاف على البطاقة التقليدية التي تصنف الاساليب. يؤسس مهندس usability مجموعة الاصناف و عدد من المواد التي ستخصص بمواضيع الأختبار الى تلك الاصناف. المهندس يستطيع أن يقارن المهام الفعلية بالمستعمل المقصود لتأكيد تلك الاصناف تجاري بديهيات المستعملين

## 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:

- اوقات تغيير سريعة وتتغير الى مواقع الويب تحدث كثيرا،

- تقدم في مستعمل وصلات تغيير مستعمل التوقعات على الانترنيت،

- حدس 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)

Examples;

a) When there are cases of (VS) structure, or full (VSO) structure foe example, the system in some cases applies the rule of Arabic parsing,

Examples:

تطرح الطبيعة الديناميه...

يصدر نسخة جديدة من الويب...

b) The modifying adjectives carry the same noun diacritieization.

Examples

- يصدر نسخة جديدة،

- الأساليب الآلية

- تدعم الاختبار الآلي البعيد والسريع،

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

Examples

- الأساليب الالية لا تستطيع ..:

- أتمت الأساليب ...

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

## Morphology

[n general, the system follows Arabic morphological rules.

Examples:

1) Words like and are formed from the root ( as ( تقيم ) and ( ) consecutively.

2) The system can generate derivations like and Others like جمع, مجموعة, مجموعات

3) The broken plural is recognized in the system أسلوب, أساليب, نموذج, نماذج and others.

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

## Lexicon

I The key word of the text is usability- It was in some cases translated BS in other cases it was transliterated as 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 testbedis transferred (one word instead of a compound term).

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

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

4) There ere key words 'Which the system failed to 'understand', and they were given differenl meaning. Example: critical in the first sentence, The word critical here means very important. The systern transferred the word as The problem here is mmnly of pragmatics nature, The

system needs to 'enjpy' world knowledge or even 'common sense' '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,

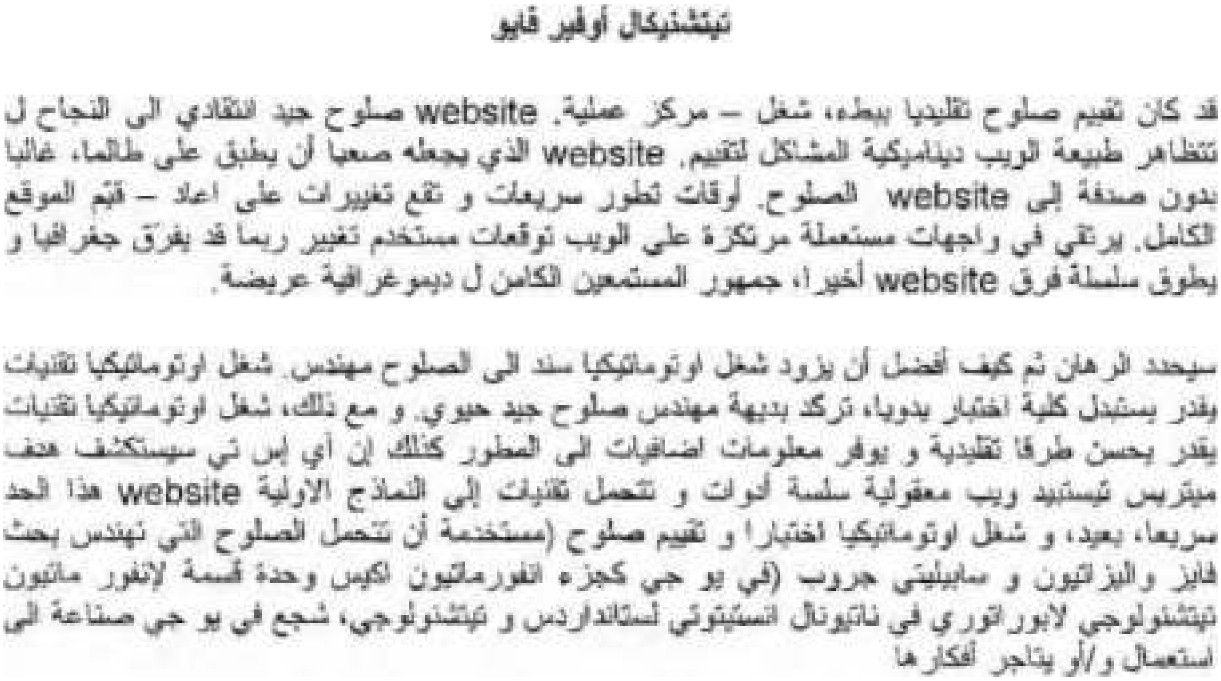
6) The technical terms of the prototype section are transliterated. This strategy can be helpful for the Arab users know the terms only in English. But on the other hand, the transliteration of terms means that Arabicization is hindered,

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 Darl*lnet categorize as "unacceptable message" in translation due to literal translation (item h),

## Al-Kati

The translated version



تيتشنيكال أوفير فايو

قد كان تقييم صلوح تقليديا بيطء، شغل – مركز عملية. website صلوح جيد انتقادي الى النجاح ل
تتظاهر طبيعة الويب ديناميكية المشاكل لتقييم. website الذي يجعله صعبا أن يطبق على طالما، غالبا
بدون صدفة إلى website الصلوح. أوقات تطور سريعات و تقع تغييرات على اعلاء – قيم الموقع
الكامل. يرتقي في واجهات مستعملة مرتكزة على الويب توقعات مستخدم تغيير ربما قد يفرّق جغرافيا و
يطوق سلسلة فرق website أخيرا، جمهور المستمعين الكامن ل ديموغرافية عريضة.

سيحدد الرهان ثم كيف أفضل أن يزود شغل اوتوماتيكيا سند الى الصلوح مهندس. شغل اوتوماتيكيا تقنيات
يقدر يستبدل كلية اختبار يدويا، تركد بديهة مهندس صلوح جيد حيوي. و مع ذلك، شغل اوتوماتيكيا تقنيات
يقدر يحسن طرقا تقليدية و يوفر معلومات اضافيات الى المطور كذلك إن أي إس تي سيستكشف هدف
ميتريس تيستيد ويب معقولية سلسة أدوات و تتحمل تقنيات إلى النماذج الاولية website هذا الحد
سريعا، بعيد، و شغل اوتوماتيكيا اختبارا و تقييم صلوح (مستخدمة أن تتحمل الصلوح التي تهندس بحث
فايز والبزائيون و سابيليتي جروب (في يو جي كجزء انفورماتيون اكيس وحدة قسمة لانفور ماتيون
تيتشنولوجي لابوراتوري في ناتيونال انستيتوتي لستاندارد س و تيتشنولوجي، شجع في يو جي صناعة الى
استعمال و/أو يتاجر أفكارها

ويب إن أي إس تي متعهدا به في 1997 استكشاف حلول الى المشاكل testbed كان ميتريس (الموصوفة فوقز خلصنا 1 اصدار (الذيتالف من اصدارات ابسط من ويبسأت، ويبكأت، و ويبفيت في يونيو/حزيران 1998. منذ ذلك الوقت، قد حسنا البرنامج الأصلي و جمعنا إف إل يو دي مع فايزفيب اوكتوبر / تشرين الاول 2002) نحرر اصدارميتريس ويب جديد كذلك مع تحسن. من فضلك أنظر) نيو صفحة و اتنا للتفاصيل.

يزود نموذج اولي من ميتريس تيستبيد – دونلواد الويب ضد صلوح web ويبسات) يراقب ستاتيك عناليزير تول الويب لغة توصيف النص المتشعب رقم) توجيهات، سواءا له يملك، لو شوط لاي إي إي إي سند توجيهات 1999-2001. يقدر يراقب webite صفحات فردية أو مزهل.

و بيكأت) يمكن كاتيجوري عناليسايز تول الويب الصلوح من ان يهندس بسرعة متشأ و يقود) إنه تغيير فوق بطاقة تقليدي يفرز تقنيات. يثبت web تحليل صنف بسيط من جانب الى أخر المهندس الصلوح شوط أصناف و عدد من المفردات الاتي سيخصصن بواسطة اختبار اخضع الى تلك الاصناف. يقدر المهندس يقارن الوجبات الحالية بالاستعمال المقصود ثم أن يتأكد أن الأصناف تباري الاصناف بديهيات مستخدمين.

## 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,

-website- صلوح جيد لتتقادى الى نجاح ...

- تتظاهر طبيعة الويب الديناميكية المشاكل لتقييم الذي يجعله صعبا بطيق ...

2) Sentences are incomplete. Examples,

- أوقات تطور سريعات وتقع تغييرات الى اعادة قيم الموقع،

- تركد بديهة مهندس صلوح جيد،

3) Punctuation is used randomly. Examples,

- قد كان تقييم صلوح تقليديا ببطء ، شغل- مركز عملية.

- وتقع تغيرات الى إعادة – قيم الموقع كامل،

4) It is impossible to apply any parsing in such sentences. However, the first sentence قد كان تقييم صلوح 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 صلوح . It seems the derivation is based on the pattern ( فعول ), 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،

استكشاف – استكشف – كشف،

أوقات - وقت ،

أشكال - شكل،

مستمعين - مستمع،

Lexicon

1) The system failed to [i] recognize' the referential meaning of some key words such as; critical

2) Transliteration strategy was used in many cases; the title Technica/ Overview was transliterated as تيتشنيكال أوفر فيو.

b) proper names are like [E]. فايزو و الزاليون و وسابيلتي جرو for Visualization and Usability Group.

c)Technical terms such as testbed ( metrics as (ميتريس).

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 mo try to post-edit it.

## 4.5.4 News Article/Political

This news article is published in the Gulf Today daily on the 22[M] of April 24, 2004. The title of the article is Arabs describe Riyadh anoeks 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 ar on board the plane wilere only English news papers are available, 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 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 ofthe security agencies in Riyadh and we express our condolences to the families of the victims and hopes for a speedy teeovery 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 Kaftaror also denounced the bombing as "barbarous and inhuman 'i.

"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 e public message .

Kuwaiti Prime Minister Sheikh Sabah al-Ahmed A Sabah also denounced the "criminal explosion' during a telephone cali 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 *gainst the precepts of
Islam and human and moral values".

Saudi LIS 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 Condoleezza Rice, pledged his nation would "fight thetn (terrorists) hard*' there will be no compromise .

**Al-Wafi**

**The translated version**

**يصفُ العرب هجمات الرياض كبربرية**

الرياض: أدانتْ الدّول العربية يوم الخميس كعمل "إجرامي"، عملية تفجير السيارة الإنتحارية في العاصمة السّعودية التي قتلتْ على الأقل أربعة أشخاص وجرحتْ 145، وقالتْ الهجوم إنتهكَ مبادئ إسلامية.

"ندينُ هذا المجرم والعمل الإرهابي ضدّ بناية وكالاتِ الأمن في الرياض ونحن نبدي تعازينا إلى عوائل الضحايا والأمال لشفاء عاجل لجرحى، "الرئيس السّوري بشار الأسد قال في رسالةِ إلى الملك فهد عاهل العربية السعودية، طبقاً لوكالةِ أنباء صنعاء.

السّلطة الإسلامية الأعلى في سوريا، الشّيخ أحمد كافتارو، شجبَ القصف أيضاً كـ"بربري ولا إنساني."

"أولئك الذين يرتكبونَ الأفعالَ البربرية ولا إنسانية بعيداً جداً عنْ روح الإسلام. بهذه الطريقة يخدمونَ أعداء (عربي) أمّة، "المفتي قالَ في رسالةِ عامّة.

شجبَ رئيسُ الوزراء الكويتي الشّيخة صباح الأحمد الصباح أيضاً "إنفجارَ إجرامي" أثناء مكالمة هاتفية في وقت متأخر من يوم الأربعاء مع ولي العهد السّعودي الأمير عبد الله، حسام وكالةِ الأنباء المحلي السّعودي الرّسمي ذكرَ.

في الدوحة، ناطق بلسان وزارة الخارجية قطري قالَ "مثل هذه الأعمال الإجرامية تضيرُ ضدّ تصالح الاسلام والقيم الإنسانية والأخلاقية."

السّفيرُ الأمريكي السّعودي الأمير بندر بن سولتان سيد في واشنطن التي الهجوم إستهدفَ "الشّعب السّعودي والعائلة المالكة ومسؤولين الحكومة. . . كلّ المواطنون السّعوديون." بندر، الذي تكلّمَ يوم الأربعاء بعد إجتماع مُستشار الأمن القومي الأمريكي كوندوليزا رايس، تعهّدَ بأنّه "يحاربُهم (إرهابيين) بشدّة "لنْ يكون هناك مساومة.

## 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 cho ice of words can be more accuratc, 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 (عربي) أما The correct structure is يخدمون أعداء الأمة العربية

2) There is one case of inconsistency between conjunction particies,• نحن أعداء الأمة ... ونبدي تعازينا

والأمال لشفاء عاجل لجرحى ....The last part ShOLlldbe ونأمل الشفاء العاجل للجرحى

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

Example;

a) The title reads يصف العرب هجمات الرياض كبربرية instead of بأنها بربرية.

The text says "شجب القصف أيضا له "بربري ولا انساني". If a phrase like Or ... معتبرا اياه عملا the translation will be perfect, This structure is repeated many times.

b) The last sentence reads تعهد بأمته" يحاربهم ( أرهابيين) بشدة" "لن تكون هناك مساومة" The right structure, I believe, is •تعهد بأن تحارب أمته الارهابيين بشدة وألا تكون هناك.

## Morphology

The system has fully succeeded to comply with the morphological rubes of Arabic language, NO

frilure is noticed.

## Lexicon

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

Examples:

a) .. وقالت الهجوم لتهك, a better choice is--واعتبرت الهجوم .

b) .. هذه الأعمال الإجرامية تصير ضد ..,‖better is... ضد

2) The system made a very critical mistake when it confused the abbreviations of the Saudi News Agency SPA asحسام وكالة الأنباء المعلني الرسمي,

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

The last two points indicate the shortage ofdatabase for Enth 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 ach ieve this result in less than two seconds as the system did.

## Al-Kafi

### The translated version

يصف العرب اعتداءات الرياض كبربريون

الرياض: تعمل دول عربية في الخميس المدان كـ"مجرم" مثل سيارة مفخخة الانتحار في العاصمة السعودية التي قتلت على الأقل أربعة ناسة وتجريح 145، وقالت الاعتداء انتهكت اسلامي مبدأ تدين هذا المجرم وإرهابي لم ضد بناء وكالات الأمن في الرياض وعبر عن تعازينا الى عائلات "الضحايا ونتمنى مسرع استرداد للمجروح،" قال بريسيدنت بشار ال أسد سوري في رسالة الى كنغ فهد للعربية السعودية، استنادا الى وكالة أنباء صنعاء.

أعلى السلطة في سوريا، الشيخ أحمد كافتارو، أيضا نقد القذف بالقنابل كـ"بربري ووحشي". الذي الذي يضع أعمالا وحشية وبربرية هو جدا بعيد عن روح الاسلام. بهذه الطريقة يخدمون "أعداء (عربي) الأمة،" قال المفتي في رسالة حكومية.

كويتي الانفجار " "الاجرامي أيضا أثناء al-ahmed نقد بريمي منايز تير الشيخ سباه ال سباه تليفون نداء الأربعاء المتأخر مع كرون بريانس أبدولاه سعودي، حضر إس بي أي وكالة الأنباء السعودية الرسمية.

في الدوحة، قال ناطق رسمي وزارة الخارجية لقطاري هكذا "تعارض أعمال اجرامية سلوكيات الاسلام وقيم أخلاقية وانسانية".

قال بريانس بندر بين سولتان سفير الولايات المتحدة السعوديات في واشنطن التي هدف الاعتداء الى السعودي ناس والملكي عائلة وموظف الحكومة ... كل المواطنين السعوديين." بندر، من تكلم في "الأربعاء بعد أن الذي قابل كوندوليزا ريس مستشار أمن مواطن الولايات المتحدة، تعهد أمته قد تقاتلهم (إرهابيون) بكد "لن تكون هناك تسوية".

## 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 strueture of the text is odd to Arabic system.

## Morphology

Morphology is usuaily the successful part of both systems, and it is the only successful part or Al-Kafi, Derivational and inflectionai fbrmation of words were done according to the morphological rules of Arabic. However, there are very few miscakes,,

Examples:

1)The formation of the broken piura2 of is oddly formed as

2) The noun is again oddly formed as in...وتقضي مسرع استرداد.

## Lexicon

Opposite to morphology, lexicon is the part cithe 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 ognize'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 every day news in newspapers, TVs, and radios.

Examples;

I ) Presidcnt Bashar Al-Assad is transliteratedas بريسيدينت بلشار ال اسلد .
2) Prime Minister Sheik Sabah Al-Ahmed is trensferred as برايمي مينيزتر الشيخ ساباه ال ساباه aiAhmed.

[n other instance, the telephme call is transferred as تلد

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

It seems that Al.Kaft 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-Kari system developers need to improve and expand the scope of the syntactic analyzer especially in terms of sentejtce 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 0?texts becomes clearer, In regard to mopphology, the system, it seems has employed good morphological analyzers, Although morphological analysis is not an easy process, it proved successful contrary

t. 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-Kan needs to re-evaluate its whole system of translatvon. The standard of translation, as it appeared from the translation of three different types of texts ig very poor, unnatural, unreadable and inadequate to the users' needs.

## 4.6 Conclusion

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

l) 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 or a TL lexicon and then restructured before transfOrrning tha sentences according to the syntax Of TL (Hutchins, 1986). However, the pn:vious corpora analysis demonstrates that Al-Wafi used a word fot word and the literal translation strategies in their translations.

Al-Kafi failed eyen to foliow the simple straight forward translating strategy; word-for-word, Al-Kaffs output is merely a combination of words put together randomly without B strategy or 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

con fused.

3) With regard to symtax, 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 ofanalysis out of text, i.e as individual cases),

4) Both systems, Al-Kati and Al-Wafi, demonstrated very goad ability to analyze and generate forms of Arabic words according to the rules and structures of Arabic morphological rules (such
as lerivational 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, tlte translation of the medicine prescription done by Al-Wafi is readable to certain extent. The reader can get the gist out of 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 Inedical 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 tanguage of the ST was not straight forward because the system cou)d not "realize the meaning very well. However 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 may be due to the faet 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 far Al•Kafi, the translation is poor and unreadable and hence the meaning is almost lost.

Post-editing does not woqk here,

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

l) As is mentioned in (item be 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 eontext, 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-Kati).

2} "ITIe two translation strategies used by both Al-Wafi and Al.Kafi, when a strategy is availables are clear representations of Nida's examples of translating improvement from word-for-word strategy to literay translation (item d, p 97). Nida's example follows

Greek
3 6
egenetø antht•öpos, apesaimenos para onoma auto löannés Uteral
transfer (stage

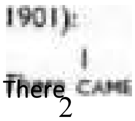
8
beciméhappened man, sent from

al trânsfer
I
CAME/WAS

name to-him John Mlhitnal
(stage 2):

2 5 6
There o man. sent from Cod, WHOSE name was John

Literary grarvfer (stage 3, example *ken front the
American Stondørd Version.


1901):
I
There CAME

There
2 s 67
o mane sent from Cod. WHOSE name John

(example eken from Ph'ilfips New Testament in *Modern English 195B [i]):

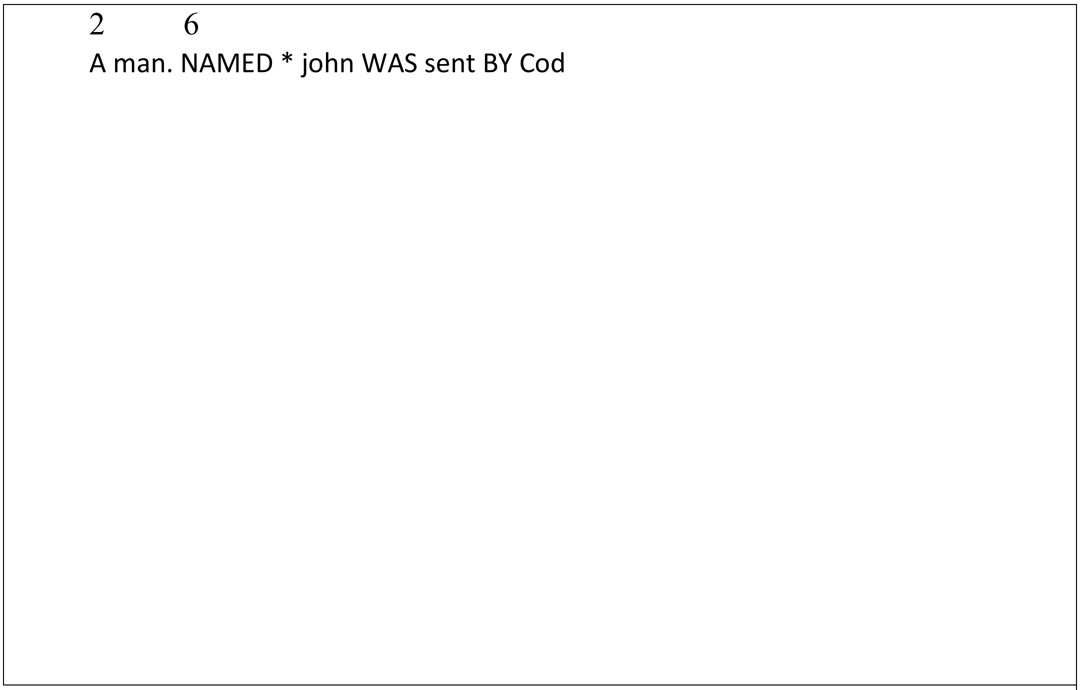


Figure (10): Nida i s model of translation improvement.

3) According to Nida's Tour basic requiæments' for the success of translation (item all four metrics are lcst large the translation outpot of Al-Kafi. In Al-Wafi, translation of some eases makes setis.e, the second medical text and the political text have instances of natural and easy of expression. they almost sitnilar response since the message is clear, 4)Vinay's and Darhe!net's indicators of the unacceptable message in translation (item h, p, 98) apply to the translation output of Al.KBfi. As for Al-Wafi, this applicable where the strict wordfor-word strategy is followed

## 4.7 Recommendations

I) 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 applicabiiity 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 suppolted 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' (he pragmatic meanings, Al strategies should be employed. Jn addition, rich databasesand encyclopedia and the adoption of examplebased 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, cither pre-editing or post-editing during the process oftranslation,

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 requines exp:essions fbr 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 translationr

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 Of 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 infornation 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 internetional communication and intellse 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. A rtother related explanation for changes in perspectives on machine translation is the information explosion, On a more practicaj level, theåe are also political factors in the search for good quality MT, In Europe, multilingualism is fact of life, which makes translation necessary for communication. However, trenslation is time consuming and continues to be expensive, so

MT could be a financial blessing-

14t

Approaches in MT are very diversified. Some researchers see MT as a means of demonstrat in their theories, with their measure cf success based on whether 0T not the system is an accurate model of human mind or simply a 'pure' theory. Other researchers concentrtte only on applying formulas lacking theoretical grounding. In facts research in MT is still, above all, experimental but guided by solid theoretical foundations. Its sole r*rformanee criterion is to obtain results fora well-defined need. There is no global solutiont however, for every translation need there is an adapted MT solution that considers the expected results and constraints on resomvcs, cost and time.

Machine translation technologies pose a range of theoretical software, hardware and even sccioh)gical 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 expenise and resources.

The level of comp lexities of a MT system depends cm the relative relationship between syntactic levels and other linguistic aspects of the source and target languages. In a direct translation stralew, 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 end also on morphological analysis. In the case of the transfer method, the SL text is analyzed and transferred into an interrnediate language called a meta-]anguage with the help of a TL lexicon and then restructured before transforming the sentences according to the syntax of T L. tn the case of Interlingua strategy, an intermediate or univetsal language 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 :anguage 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 systerm 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 syntacties. It arises mostly due to the inadequate details oi semantic representation and inefficient techniques adopted to represent the ambiguous situations and contextual variations, The most complex NL problems as related to MT ate symtaetie ambiguity, lexical and semantic ambiguities and idiomatic expressions, pragmatics or language in context. ellipsis. substitution and anaphofic references.

Forwnately, 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 decisi.ons and resolves persisting ambiguities sinw no program is able to integrate sufficient world knowledge and common sense so as to automatically resolve ail 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 tm 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 current!y on the Intemet 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

Sinec 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 2anguage engineering and renewal of Arabic by initiating a fresh formulation of its grammatical rules to meet the requirements of computational processing, It is also sential toconsolidate and etihanee glossaries of specialized terminology and thesauruses

Unfortwnatcly, Arab countries are still lagging behind because there is a lack of interest from the Arab financial sector in information ptojects, were feasibility studies are normally undertaken on a purely economtc 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. tn 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 lanyage processing as a natural language, to ocknowledge 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 bath 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 severål 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 (morpholog, syntax and semantics),

Arabic, as a Semitic language diftérs from European languages morphologically, syntactically and semantically. Most words are formed from a tii.lateral roots which falls imo specific patterns: a key morphological feature. Though there has been much interest recently in handling morphologically rich inflectional languages such as Arabic, the Ambic language is somewhat difficultto deal with due to its right to leh 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 every day 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 Acahic lies in the fields ofdiscourse and pragmatic.

The future of MT is bright if wc remain realistic. To obtain a tran51ation of suitable quality, hyhrid and innovative approaches must be relied upon, This includes using large and comprehensive dictionaries, a wide range or 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 parallelcorpora and statistical systems provide possible solutions for today. Future improvements in computer hardware and software and in ;anguage technology and engineering may create machine that can replace human translators, This is a dream not to be lealized 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

Ali N. (2003), The second wave of Arabic Natural Language Processinv A content Perspective, Retrieved January 10, 2004 from wvwæsewa.com

Ali N, (2003), Standardization related to Arabic language use in ICT,, Retrieved on February 2, 2004 from www.eswa org.Isl•vvsishneetingsJ3-5 June.

Ali N. Machine Translation: A contrastive linguistic perspective. Retrieved on January 8, 2004, from www.unesco.org.

A ustermuhl F. (2001). Electronic toolsjor transtøjors. Manchester: Stjercrne,

Bass S. (1999), Machine Vsr I-luman translation, Retrieved on December 9, 2003, from www.advancedlanguage.com.

Beesley K. (2001 Finite-state Morphological analysis and generation ofArabic at Xerox Research; status and plans. Retrie•ged on January 2004, from www.elsne.t.oqg.

Belis A, An Experiment in Comparative evaluation; Humans vs, Computers, Retrieved On Deccmbcr 15, 2003. from www.amtaweb.org/sumit/mtsummit/finalpapers/.

Brace C. (1990). Japanese to English Machine translation: A Repott from a Symposium. Retrieved on February I 20049 from http:.<sup>i</sup>/books.nap.edwbook.

Carbor,ell L, Rich E., Johnson D., Tomita Vaseoneellos MI, Wilk's Y. (1992b Machine Translation in Japan Retrieved on December 29, 2003, from www.wtec.org,

Chalabi A. (1997). Arabic language software issues: Its development, technology and industry. Retrieved on December 29, 20039 from www.georgetown.edu/research/arabtech

Diab H. (2003). Standardization Related to Arabic Language Use in ICT. Retrieved an Februaj•y 28, 2004. from wwwesewa.ocg.lb/meetings,

Faiq S, (2000). Arabic Translation: A glorious past but a meek present. In Marilyn Gaddis Rose (Ed.). Beyond the wesjent rradirion; Iranslaåonperspective M, 'New York'. State University of New York at Binghamton.

Flanagan, M. (2002). Systran and the reinvention of MT. Retrieved OTI December 12s 2003, from www„sY5trN.D5QLcom.

Foster Gandrabur S.. Langleis Plemondon Pm, Russel G.. Simard M. Statistical Machine Translatiom Rapid Development with Limited Resources, Retrieved on December 8, 2003, from www.amtaweb.org/summit/mtsummit/finalpapers.

Guidere M. Toward Corpus-based machine translation. Retrieved on December 23, 2003, from wvv•€v linguistic.0[.L'.

Hatim B., & Mason l . (1 990), Discourse and 'he Translator. London: Longman,

Hutchins J, (1979), Linguistic Models in Machine translation from: CIEA papers in linguistics 9. Retrieved on January 3, 2004, from http://ourworld.compuseo•c.com,

Hutchins J, (1986). Machine Translation: Past, Present, Future, U.K.: Ellis Homoad & New York' Halsted Press.

Hutchins & Somers, H.. (1992), An Jnrroduction to Machine Translation. London: Academic Press.

Hutchins J. (1999). The Development end use of machine translation 5ystems and computer-based translation too}s. England; University of East Ag]ia, Retrieved on December 8, 2003: from ourworld,compuserve.com

Hutchins J, (2001). Machine Translation and Human Translation; in competition or in complementation. In Blekhman, (Ed), Speeiaf Theme Issue Machine Translation. Norwich; International laurna] of Translation. Retrieved on March 20, 2004, from http:.nourworld.ccmpusewe„eom„

Hutchins J, The history of machine translation. Retrieved on December 3 2003 from cutworld.compuserve.com.

Hutchins J. "'e state o/
March I 2004 from
machâne achine translation in in Europe andfuture pyospeetf. Retrieved
www.hltcentral.org. on

Hutchins J. The state of machine translation in Europe. Retrieved on December IOS 2003, from htcpzf/ourwodd.compuserve.com„

Mariani J. Are we losing ground to the LS? Retrieved on March 3, 2004* from wy..u.httcentczl.org.

Marzouki M. (2002). The Golden Wan & Artificial Intelligence. Retrieved on October 12, 2003 from www-alriy.adb:u.com,,

Munday (2001 inrroducing Translation Sffådies. London and New York; Routledge,

Napier M. (2000) The Soldiers are in the coffee. An introduction to machine translation. Retrieved or, October 12, 2003, from www.cylüypt$-int.ocg-

Newmark P. 1988). Approaches fo Tyuns/ajion, Prentice Hail International.

Osborn D, (2003). Arabic Language Processing. Retrieved on December 2003, from http;/'l ists.kabbissa.org/[ists/atchives.

Raddawi± R, (2000). Computer Enderdisposal of Translator- King Saud University. Riyadh.

Raddawi, R. (2004). Machine Translation; Globalization and Localization. A public lecture at the American University of Sharjah .

Rao O. (2001). Machine Translation in India; A brief survey. Retrieved on December 31. 2003, from www.eldAfr&L.i'proiisailQfsggLla.JOl.

Sabah S. Future of Arabic language teehnoiogy. Retrieved on December ] 0, 2003, from www.ccse.kfupm.edu.sa.

Seasly Machine Translation; A survey of approaches. Retrieved on March 2004, from P.yywuetsongLumich.ed4

Srikanth RP & Dheeraj. Machine translation set foc quantum leap in India, Retrieved on www.express.computerline.com.

United Nations Development Program: Arab Fund for Economic and Social Development. (2002). Arab Human Report, Creating Opportunities for Future Generations. Amman.' Natiaqul Press.

United Nations Development Program: Arab Fund for Economic and Social Development(2003b Arab Human Report, Building a Knowledge Society. Ammam National Press.

[Jszkoreit I-I, What is Computationai Linguistics? Retrieved on March 2, 2004, from www É9J.i.uni.sb.dg.•:

Uszkoreit H. Statistical Modeling: Oveqview, Retrieved on February 25, 2004, from http://cslu.cse.ogi.edu/hltsurvey/ch//node3.

Viney J & Darbe)net J. ( 1958). Stylisiique Comperee du Francois el de LÄonglais, Merkode de Traducnoø, Paris: Didier,

Yaseea M, Haddad B. Towards Understanding Arabic: A logical Approach for Semantic Representation. Retrieved on December 29, 2003, from Eyyvtelsnet„or.g.

## Arabic Publications:

- محمد حناش (2004). ماذا يقدم الحاسوب لمعالجة العربية بوصفها لغة طبيعية . الامارات العربية المتحدة: جريدة الخليج، 12 •e..g.l•.h يناير 2004، العدد 9003.

- نبيل علي (1989). اللغة العربية والحاسوب . القاهرة: تعريب

- نبيل علي (1994). العرب وعصر المعلومات . الكويت: سلسلة عالم المعرفة – المجلس الوطني للثقافة والفنون والآداب.

- الترجمة في الوطن العربي: نحو انشاء مؤسسة عربية للترجمة (200). بحوث ومناقشات الندوة الفكرية/ مركز دراسات الوحدة العربية. بيروت: مركز دراسات الوحدة العربية.