CHAPTER ONE

RAISON D’ETRE AND NLP LITERATURE REVIEW

a. RAISON D’ETRE OF THE THESIS

With advances achieved in Linguistics and Computer Sciences, the scope of language-automation in general and translation-automation in particular has broadened to become all the more effective and efficient. Techniques in NLP (Natural Language Processing) have per force benefitted from the study of language. With increasing awareness of the importance language-automation, several scholars and theoreticians have exerted intense and sustained efforts to build and theorize the phenomenon of natural language engineering and translation-engineering. NLP applications in the English language have already gone a long way in this regard. Other European and Asian languages have also made sizeable contributions in this field.

The case of Arabic however is different. NLP is still in its nascent stages as far as the Arabic language is concerned. The development of NLP and its applications for Arabic require availability of different techniques to process M(odern) S(standard) Arabic texts. These are not routinely available. Attempts have been made to build Arabic language applications, but the fact is that the output of these attempts has been less than desirable, owing to difficulties arising from the scantiness of electronic texts, the nature of the Arabic language, its writing system, and the lack of reliable data for experimentation.

Beyond the traditional challenges of natural language processing and natural language engineering for English, there are unique complexities for the Arabic language for processing and translation, some of which are listed below:
1. Lack of diacritical marks: Arabic texts are usually presented without short vowels, which are indicated in Arabic by diacritical marks placed above or below the characters.

2. Free word order: The Arabic language has considerable freedom in word order, where in typical cases sentence-constituents can be swapped without affecting structure or meaning.

3. Scarce punctuation: In regular usage, the Arabic language punctuation system is not as widely developed as the English counterpart, but progress is being achieved in his regard. Although the Arabic language has punctuation marks, written Arabic employs these punctuation marks in a way that differs from the English usage. For example, a full paragraph in Arabic could contain only one full-stop, with clauses separated by the use of commas. Therefore, paragraph-long sentences that require automatic sentence segmentation prior to analysis are common. Other issues that add to the complexity of the Arabic language include right-to-left direction of the text, inflectional writing, significant characters, and more.

There is some evidence that statistical NLP techniques for information retrieval (IR) on European languages do not transfer well to Arabic because of the nature of the language and its writing system (Yahya 1989, Hmeidi et al. 1997, De Roeck and Al-Fares 2000). Experimentation in Arabic language environments has been relatively new and significantly limited, compared with work on English and other European or Asian languages.

The thesis sheds light on different facets of NLP and its applications and investigates the existing Machine Translation techniques from English to Arabic, their usefulness and effectiveness. Since they are still in their inceptive stages, these

2
machine translation systems have to be able to overcome the limitations and difficulties and new techniques developed to improve upon the current status of Machine Translation in Arabic.

The thesis goes some distance in addressing this part of NLP and MT in Arabic. The thesis is also an affirmation of the belief that rule-based MT is the right way forward simply because language is NOT a statistical or probabilistic or stochastic phenomenon. This idea is filled with empirical content by contrasting the fragment of MT software we have put together with outputs from Google Translate, System and other MT programmes available. Output from statistical MT programmes is notoriously non-rigorous.

NLP techniques are the outcomes of the interface between computer and language, the application of which results in finding out new evidence to describe language and its properties from different perspectives. The marriage of Linguistics and Engineering has given birth to techniques like Information Retrieval, Information Extraction, Statistical Frequency Count, Lexical Collection, Key-Word-in-Context, Concordance, Local Word Grouping, part-of-Speech Tagging, Morphological Processing, Lemmatisation, Text Annotation, Parsing, shallow and deep, and morphological analyser and generator and so on.

All traditional semitic-language dictionaries and most modern ones are arranged by root. Instead of listing alphabetic entries, these dictionaries arrange words under entries of the roots that produce them and the roots arranged alphabetically. To look up a specific word, the user has to have enough knowledge to extract the root and then locate its entry. It is as though words like ‘ascribe’, ‘describe’, ‘subscribe’, ‘circumscribe’, ‘proscribe’, ‘prescribe’, ‘inscribe’, were listed in an English dictionary under the Latin root scriber that describes the basic idea of writing/drawing (De
The difference is that the words grouped under an Arabic root can be analyzed down to the letters of a root and the predefined morphological patterns that created them.

One of the standard Arabic lexicons, viz. *Tongue of the Arabs*, lists 6,350 triliteral roots and 2,500 quadrilateral ones. Out of these, only about 1200 are still used in modern Arabic vocabulary (Hegazi & Elsharkawi 1985), and the great majority of words can be analyzed down to trilateral roots consisting of three consonants or radicals. Words constructed from the same root constitute what is traditionally called a morpho-semantic field, where semantic attributes are assigned through patterns governed by morphological rules. The meaning that is inherent in the root is shared by all words in this field. However, the patterns that produce these words make them semantically distinguished (Rafea & Shaalan 1993). A similar process can be noticed in English if we look at ‘necessitate’, ‘necessary’, ‘unnecessary’ and ‘necessarily’. While all four words share the basic meaning that is inherent in *neces*- (Latin *need*), they convey different semantic messages: necessitate (to produce the need), necessary (needed), unnecessary (not needed), and necessarily (in need condition/mode). We could say that adding –*itate* to the root creates the verb ‘necessitate’, -*ary* the adjective ‘necessary, and so on.

McEnery and Wilson (2001) describe a modern corpus as any collection of more than one text with four main characteristics: sampling and representativeness, finite-sized, machine-readable form, and status as standard reference. But because Arabic corpora are hard to find, no extensive findings have been published to confirm that Arabic is problematic for standard techniques. It is therefore evident that although the techniques used in building the corpora of, say, the English language, are useful as
guidelines on how to do the same for the Arabic language, putting them into practice with reference to Arabic cannot claim to yield the desired results.

**SCOPE**

The present study is a language-engineering probe into, or in particular, an MT investigation into Arabic. There have been some Arabic-English MT systems built but none in the other direction viz English to Arabic. This study thus pioneers English-Arabic MT as it does rule-based MT in Arabic. The pioneering and path breaking nature of this study becomes clear in the part of this study where we contrast our outputs with the output in translation sites like Google Translate and Systran.

**SOP (STATEMENT OF PURPOSE)**

The present thesis will seek to do the following:

1. Surveying and taking stock of the NLP and the MT scenarios in Arabic.
2. To point out the feasibility of developing and using corpora for the Arabic language, and highlight its usefulness not only in language studies but also in other fields, such as computer sciences, hence the electronic literature of all the fields of knowledge written in Arabic. That is to say, how MT studies of the Arabic-English pair can benefit from such a step.
3. To discuss the techniques used in building rule-based MT systems, their advantages and disadvantages.
4. To build a rule-based MT system in a small area of grammar. We choose the realm of imperatives to do this, as an illustrative demonstration.
5. To propose solutions for the difficulties and to develop techniques to further develop the existing state of MT with reference to English-Arabic translation.
METHODOLOGY

The methodology followed in this thesis is two-fold: theoretical and practical. The theoretical side reviews and does a (de)constructive survey of the literature of translation automation studies, and Computational Linguistics. The practical application side comprises an MT package to translate from English to Arabic.

When first introduced, transliteration of the Arabic word will be accompanied by its English equivalent translation in the following format: Arabic transliteration, English translation.

The transliteration scheme used is a proposed one to enable mapping the letters of both languages for accurate output and is detailed later in Table 3.1. If the same word is used again, only the translation will be used and will be placed in quotes.
1.1 INTRODUCTION

Natural Language Processing (NLP) is computer-processing of natural language to simulate human linguistic behavior. *It comprises a whole theoretically motivated gamut or range of computational techniques for analysis and generating naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.*

When a ‘range of computational techniques’ is taken into consideration it means there are many methods or systems to choose from to accomplish a given task for a particular type of language analysis. Texts that occur naturally can be of any language, mode, genre, etc either in oral or written form where the only criterion is that it is a language spoken by people to express oneself and communicate with one another. The text for analysis is a very natural one and not a constructed one for the sake of analysis.

As there are multiple types of language processing understood to be at work when people talk to each other, the idea ‘levels of linguistic analysis’ refers to this. Though it is thought that people use all the different levels, NLP uses the various combinations of these levels of linguistic analysis and this difference is seen among the different NLP applications. So those who are not specialists find it difficult to understand these differences and the question of weak NLP and strong NLP arises among the applications.

NLP is considered as a discipline within Artificial Intelligence (AI) as NLP strives for ‘human like language processing’ even though NLP depends on many other disciplines. Except for AI researchers NLP is not considered as a goal in itself or
a goal of itself as other researchers treat NLP as an application to accomplish a particular task.

1.2 AIM

The basic aim of NLP is to achieve human like language processing system. A full NLP system would be able to:

- paraphrase an input text
- translate the text into another language
- answer questions about the contents of the text
- draw inferences from the text

Though many NLP applications are able to meet the first three points, it is a fact that NLP systems cannot draw inferences from the text, which still remains an elusive goal of NLP. As research fields, NLP and machine learning still have such a long way to go, and this makes thorough data analysis using algorithms not nearly as accurate as we would like.

The reality of the situation is that human-beings are still more reliable than computer-driven alternatives. It needs to be said however that it is not clear why human behavior cannot be simulated or replicated in toto, given that all human behavior is motivated one way or the other. Linguistic creativity for example can not really be discontinuous with what already exists and can therefore be machine-learnt. Unsimulable linguistic creativity is a small, even if significant, part.

1.3 HISTORY OF NLP

NLP has a mix of lineages like many modern disciplines and NL is influenced by one or another discipline. The main contributor is Linguistics, which studies
natural language in all its facets. Another contributor is Computer Science and this helps in developing the data representation internally. Cognitive Psychology is considered as the window to human cognitive process and this discipline is also used extensively in research to develop human like language processing modules.

For several decades research has been going on in Natural Language Processing almost dating back to the 40s. The first computer-based application related to Natural Language Processing was Machine Translation. Warren Weaver and Andrew D. Booth started one of the first MT projects in 1947 on computer translation based on breaking enemy codes during the Second World War and it was generally accepted that it was Weaver’s memorandum of 1949 that brought the idea of MT to public notice and further inspired many other people to work on various projects related to this field. He suggested using ideas from cryptography and information theory for language translation. Research began at various research institutions in the United States within a few years.

In the initial years, researchers assumed that MT had only to take into consideration the differences that obtained between languages in their vocabulary and word orders. This made researchers develop systems that used the simple dictionary look-up for the correct word usage and word order rules of the target language. But things like ambiguity, lexical and structural, and rules were not taken into consideration at all during this period and this naturally produced poor results. The apparent failure made researchers realize that the task was a lot harder than anticipated, and they needed a more adequate theory of language. However starting from 1957 (the year of publication of Syntactic Structures), introducing the idea of generative grammar, there has been a steady enhancement of our insight into natural language. And the field gained better insight into whether and how mainstream
Linguistics could help MT. during this time, a number of NLP application areas began to emerge, such as speech recognition and speech synthesis. The language processing community and the speech community then were split into two groups with the language processing community dominated by the theoretical perspective of generative grammar and the speech community dominated by statistical information theory.

Researchers were over enthusiastic in the 50s about the various developments in syntactic theories of language and the general belief was that it was possible to create a high quality translation system that could produce results indistinguishable from those of human translation and such systems would be in vogue in a few years’ time.

The ALPAC (Automatic Language Processing Advisory Council), US, published a report in 1966 about the inadequacies that existed in the systems and the overenthusiasm shown in the development and research of NLP and ALPAC concluded in the report that MT was not achievable anytime soon and recommended not to fund MT research. All research work ground to a halt soon after this report was published and most of the other work in the related areas of NLP was also brought to a halt in the United States.

There was substantial decrease in NLP work after the ALPAC report but there was significant development in theoretical issues and the development of prototype systems. In the early 80s, researchers started re-examining non-symbolic approaches that had lost popularity in the early days. By the end of 1980s, symbolic approaches had been used to address many significant problems in NLP and statistical approaches were shown to be complementary in many respects to symbolic approaches.
In the last twenty years, the field has been growing very fast. This can be attributed to many factors:

a) increased availability of large amounts of electronic text
b) technological improvement increasing the speed and memory of computers
c) the development of the Internet.

Statistical approaches succeeded in dealing with many generic problems in Computational Linguistics such as part-of-speech identification, word sense disambiguation, etc, and have become standard throughout NLP. NLP researchers are now developing next generation NLP systems that deal reasonably well with general texts and account for a good portion of the variability and ambiguity of language.

1.4 RECENT DEVELOPMENT

There are a number of future applications of NLP of which those currently under development are as follows:

- *Conversational systems:* A system developed by the University of Colorado at Boulder which is designed to assist with inquiries related to airlines schedules, hotel reservations, the times of movies and their locations, or sports scores (Andolsen 2002). The huge challenge in this speech recognition system was, and which still remains to be so, is proper recognition of what is being spoken by a wide variety of people with differing vocabularies and accents.

- Development of a system where a computer would be able to read a book, store the information about the book, and then answer questions about the boo. These types of system would be dealing with an advanced type of autoindexing.
• *Artificial Natural Networks*: DolphinSearch technology is one of the interesting products now being introduced in the market. In a similar manner where dolphins recognize the characteristics of an object and learn about the object using sonar waves, this approach relates words to one another so that, in ambiguous situations, their grammatical role becomes evident. For example, the word “can” can be either a noun or a modal verb. By analyzing the words around it, the system is able to determine whether it is being used as either a noun or a modal verb.

• Microsoft MindNet: This is a combination of an extensive database and algorithms that can define relationships. In this project they attempt to use dictionaries in seven languages and a variety of encyclopedias to create a system that recognizes relationships between simple words (from the dictionaries) and phrases or sentences (from the encyclopedias). MindNet also looks like a powerful tool for machine translation. MindNet creates separate conceptual webs for English and another language, Spanish, for example, and then bring into line the Webs so that the English logical forms match their Spanish equivalents and then this system annotates these matched logical forms with data from the English-Spanish translator memory, so that translation can proceed smoothly in either direction (Waldrop 2001).

• Medication Assistant: This system is a medical system, which models the effects of therapy on patients with cardiovascular and other medical conditions. Prolog programming language is used to control NLP links with hierarchically linked data and grammatically correct texts (Temiroff et al. 2001).
Chatterbots: The new generation of chatterbots is being constantly developed. Chatterbots use natural language processing to simulate conversations with users. Web sites are beginning to install chatterbots as Web guides and customers service agents (Anonymous 2001).

1.5 CLASSIFICATION

NLP consists of two distinct categories namely language processing and language generation. Language processing refers to the analysis of language for the purpose of producing a meaningful representation, while language generation refers to the production of language from a preexisting representation.

1.5a Natural Language Generator:

The role of Natural Language Processing is equivalent to the role of reader/listener while the role of Natural Language Generation is that of the writer/speaker. Both these categories use a lot of theory and technology but Natural Language Generation requires a planning capability apart from theory and technology. That is, the generation system requires a design or model of the goal of the interaction in order to decide what the system should generate at each point in an interaction.

Wallis and Shortliffe (1985) describe one of the first NLG systems that consults a user model. The need to model two aspects of a user to generate appropriate explanations: his/her expertise in the subject matter and his/her preferences regarding the level of detail of an explanation was recognized by Wallis and Shortliffe. Their model consisted of only a single number to represent each of these aspects and this was not considered a refined system. So this showed that more sophisticated user models were required in order to enable NLG systems to generate
appropriate and relevant discourse. Researchers now consider NLG systems that generate discourse that incorporates features like presenting different types of information based on the audience’ preferences or perceived expertise, addressing user’s likely misconceptions and inferences and discuss the user models that support the requirements of these systems.

**One Dimensional User Model:**

There are NLG systems that consult user models representing a single aspect of a user, where this aspect is influenced by the type of discourse being generated. For instance, systems that generated concept descriptions consulted a model of a user’s expertise or interests (e.g. Paris 1989; Tattersall 1992; Stock et al. 1993) and systems that produced evaluative discourse used a model of a user’s preferences (e.g. Jameson 1989, Carennini and Moore 1999, Ingrid Zukerman 1999).

It was observed by Paris that a user’s level of expertise not only affects the amount of description provided in an explanation, but also the kind of information specified. This observation helped in the development of a system that fits into the type of information included in a description to a user’s level of expertise. For supporting this idea, Paris’s user model distinguished between two types of information items in the domain, basic concepts and specific artifacts known to the user. Although this model was rather unrefined in nature, it was enough to support the distinction between the two types of explanations under consideration. These strategies also determined whether comparisons were optional or mandatory, and whether known or unknown target concepts should be used in comparisons. These strategies also determined whether comparisons were optional or mandatory, and whether known or unknown target concepts should be used in comparisons.
The applications that used these models are the systems used in health care which are tailored to patients’ needs and disabilities and also systems that generate a discourse that supports an evaluative process.

**Multi-Dimensional User Model:**

The contextual information that took the form of an object perspective and highlighted the attributes in the model of a user’s beliefs which were most relevant to the user’s topic of discussion was evident in McCoy’s system (McCoy 1989). This indicated a likely misconception in the user’s mind and in the model designed by Zukerman et al. (1998) there was a similar aspect to discourse context in a system that generated arguments. The systems activated in this manner constituted focal points around which the search for an argument was conducted.

These systems consulted a multi-dimensional user model to generate extended responses to user’s plan-based queries (van Beek 1987). The responses generated by van Beek’s system showed the incompatibilities between a user’s plan and his/her goals or preferences, and provided alternatives to the user’s sub-optimal plans.

Other researchers also came up with multi-dimensional models. Multi-dimensional user models play an important role in multimedia interfaces. The input and output modalities of these interfaces may include linguistic modalities, such as text and speech, and non-linguistic modalities, such as graphics, animations and pointing. Thus, in addition to making decisions about the content of a presentation, multimedia interfaces must determine its modality, which is affected by users’ preferences and interests.
1.5b Natural Language Processing:

Levels of Natural Language Processing:

Natural language processing includes several levels or areas of language such as phonetics, phonology, lexicon, morphology, syntax, semantics and pragmatics.

a. Phonetics:

Phonetics is the dynamics of speech sounds. There are three types of phonetics

a) Acoustic Phonetics
b) Articulatory Phonetics and
c) Auditory Phonetics.

*Acoustic Phonetics* studies the way speech sounds are made (articulated) by the vocal organs, and the conversion of aerodynamic energy into acoustic energy. Aerodynamic energy is the airflow through the vocal tract. Acoustic energy is variation in the air pressure that can be represented as sound waves, which are then perceived by the human ears as sound.

*Auditory Phonetics* deals with the dynamics of how our ears perceive and interpret the sounds.

Phonology:

Phonology is a study that explains how sounds are organized and used in natural languages. It is the way raw speech sounds pattern or are structured. An important thing to note is that at the phonetic level the representations are not cognitive as it deals with the events in the outside world and not the events in the mind. Phonology and Phonetics are of course related, but have dissimilar representations. Phonological representations are cognitive, qualitative and relatively
accessible to introspection whereas phonetic representations are quantitative, non-cognitive and less accessible to introspection.

Phonology includes prosody, stress and intonation.

Example:

1. Sam told me give this to YOU.
2. Sam told ME to give this to you.

It may thus be said that phonology involves qualitative distinctions which are manipulated by syntactic rules.

In Arabic, phonological stress falls in a particular place in the text. Amman-Jordanian Arabic and Palestinian Arabic have similar stress pattern. See Table 1.1

**Table 1.1 Penultimate, Antepenultimate and Final Stress**

<table>
<thead>
<tr>
<th>(1) Penultimate Stress</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$ar$ak</td>
<td>‘he participated’</td>
</tr>
<tr>
<td>urdon</td>
<td>‘Jordan’</td>
</tr>
<tr>
<td>maktabha</td>
<td>‘her desk/office’</td>
</tr>
<tr>
<td>binsaameh</td>
<td>‘we forgive’</td>
</tr>
<tr>
<td>saamahatna</td>
<td>‘she pardoned us’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(2) Antepenultimate Stress</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>maGallamak</td>
<td>‘he didn’t teach you’</td>
</tr>
<tr>
<td>fabarada</td>
<td>‘he got cold’</td>
</tr>
<tr>
<td>?allamatake</td>
<td>‘she taught you’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(3) Final Stress</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>darast</td>
<td>“I studied”</td>
</tr>
</tbody>
</table>
b. Orthography:

Orthography maps the sounds of a language from a particular script.

There are problems with the type, style, case and punctuation used in English. When a word is in bold, italics, or underlined then the word has used type style. If a word is in lower case (word), upper (WORD) or title case (Word) then it has been using case options which are not available in the traditional Arabic script. Punctuation means how different punctuation marks (like Comma, period, colon, and semicolon) are used to demarcate sentence boundaries as well as boundaries and relations within the sentence. There are many orthographical differences between languages, particularly Arabic and English, which have to be taken into consideration by MT developers in order to make the translation as effective and efficient as possible.

To emphasise a word, English uses italics but emphasis in Arabic is indicated by a change in word order or the introduction of an emphatic word. Proper nouns start with a capital letter in English but Arabic does not have any special marking for proper nouns. There is no regular usage of punctuation marks in Arabic as they are in English. During the process of translation, the system must therefore be able to identify which punctuation marks are redundant in the other language.
The following English sentence, for example, uses punctuation marks abundantly. None of these marks is used in its Arabic equivalent.

1. This morning, the weather was very pleasant. However, the sudden rains did not let it last for long.

2. *kan aliwu jamilun fi SabAah hatA maG za#at almaTar.*

c. Lexicon:

At the lexical level both humans and NLP systems interpret the meaning of individual words. There are many types of processing that contribute to word-level understanding, the most important of these being assignment of a single part-of-speech tag to each word. In this processing, words that can function as more than one part-of-speech are assigned the most probable part-of-speech tag based on the context in which they occur.

At the lexical level words that have only one possible sense or meaning can be replaced by a semantic representation of that meaning. The nature of the representation entirely depends on the semantic theory utilized in the NLP system. As can be observed, a single lexical unit is decomposed into its more basic properties. A set of semantic primitives is used across all words, these simplified lexical representations make it possible to unify meaning across words and to produce complex interpretations, much the same as humans do.

The lexical level will require a lexicon, and the specific approach taken by an NLP system will determine whether a lexicon will be utilized and it will also determine the nature and extent of information that is encoded in the lexicon. Lexicons may be quite simple, with only the words and their part(s)-of-speech, or may be increasingly complex and contain information on the semantic class of the
word, what arguments it takes, and the semantic limitations on these arguments, definitions of the sense(s) in the semantic representation utilized in the particular system, and even the semantic field in which each sense of a polysemous word is used.

d. Morphology:

Morphological analysis is one of the most basic components in any MT system. Morphology has to do with how phonemes form from words. Analysis and generation of words is crucial in MT. Words are composed of morphemes – the smallest units of meaning. For example, the word ‘preregistration’ can be morphologically analyzed into three separate morphemes: the prefix pre, the root register, and the suffix ation. Since the meaning of each morpheme remains the same across words, humans can break down an unknown word into its constituent morphemes in order to understand its meaning. Similarly, an NLP system can recognize the meaning conveyed by each morpheme in order to gain and represent meaning. For example, adding the suffix to a verb, conveys that the action of the verb took place in the past. This is a key piece of meaning, and in fact, is frequently evidenced in a text only by the use of the –ed morpheme.

Morphemes could be free, bound, inflectional and derivational. Morphological inflections which carry syntactic meanings are:

**Person:** has three main contrasts are: first person, or the speaker; second person, or the addressee, third person, which refers to a third party.

**Number:** has four main contrasts: singular, dual, trial and plural. English makes a distinction only between singular and plural, while Arabic makes contrasts between singular, dual, and plural.
**Gender:** Its typical contrasts are: masculine, feminine, neuter, animate, and inanimate. In Arabic there is a further contrast between human and non-human objects and each has grammatical reflexes.

**Case:** This indicates the role of a participant within a phrase. Traditionally they are nominative, accusative, dative, instrumental, genitive and comitative.

**Tense:** Tense is linguisticisation of time. It basically expresses whether an action is performed in the past, present, or future time.

**Aspect:** This marks the temporal contours of an action, expressing whether an action is complete (‘He has gone’), progressive (‘He is reading’), habitual (‘He wakes up late every day’), intermittent, etc.

**Mood:** It expresses factuality, likelihood, possibility or uncertainty of the proposition.

**Voice:** It expresses the relation of the subject to the action. Its typical contracts are active ‘I teach’ and passive ‘I am taught’.

**Lemma:** A lemma is the authorized form of a lexeme, which is the set of all surface inflected forms that represent the same word. In Arabic, nouns are represented by the singular form of the noun and the lemma of a verb is the third-person singular of the past tense.

Lemmatization means the process of automatically finding the lemma of the words within a given text. A lemmatized text is a text in which every word is replaced by its lemma. Arabic lemmatization tends to be a difficult task and, currently, there are no known Arabic lemmatizers that are available for public use.

The Arabic language is well known for its morphological richness and complexity. It has been a challenge to process Arabic due to its morphology and it has always been a hard testing ground for morphological analysis technologies. Dealing
with Arabic morphology has to do with two strategies depending on the analysis and on whether they are stem-based morphologies or root-based morphologies.

Stem-based morphologies analyse Arabic at the stem level, using regular concatenation. A stem is the least marked form of a word which is not modulated, a word without suffixes, prefixes, proclitics or enclitics. In Arabic, this is usually the perfective, third person, singular verb, and in the case of nouns and adjectives, they are the singular indefinite form.

Root-based morphologies analyse Arabic words at the root level and work on roots and their patterns in addition to concatenations. A root consists of a sequence of three (rarely two or four) consonants called radicals, and the pattern is a template of vowels, or a combination of consonants and vowels, with slots into which the radicals of the root are inserted. This process of insertion is called interdigitation (Beesley 2001).

*Morphological Analysis:*

Morphological analysis means the splitting the word into its constituent morphemes. Even though the words could be analysed in different ways by different analysers, all the possible analyses are produced by most of them.

For instance, the Arabic word *wldy* ‘my son’ can be analyzed in several ways. One way is *wld + y* ‘my son’, where the first morpheme is the noun *wld* ‘son’ and the second morpheme is a first-person possession suffix. Another possible way is *w + ldy + y* ‘and with me’, where the first morpheme is the conjunction *w* ‘and’, the second morpheme is the preposition *ldy* ‘with/by’ and the third morpheme *y* is a first-person pronoun.
A single word and the number of its possible analyses has a very high proportion in Arabic like many other Semitic languages and therefore in most cases the higher processing layers are depended upon to resolve the ambiguities be it either syntactic or semantic analysis. For building a morphological analyser two common approaches are:

- Rule-based approach
- Corpus-based approach

A number of predefined rules are used in the rule-based approach which a finite-state machine represents. The Buckwalter Arabic Morphological Analyzer produces all possible Morphological Analyses of a word since it use a large lexicon with 82158 entries representing 38600 lemmas (in version 1.0), 299 possible prefixes and 618 possible suffixes. There is a predefined table with nearly 3500 entries and contains morphological rules for possible combinations of stems, prefixes and suffixes.

Corpus-based analyzers learn how to handle new inputs using a large tagged corpus, in the case of a supervised analyzer, or a large untagged corpus, in the case of an unsupervised analyzer. Hybrid analysers use template rules to find the root of a new given word, and also a large corpus for resolving ambiguities by statistical methods. The Sebawai Analyzer extracts the template rules automatically from a table that is created by another morphological analyzer. The table contains many Arabic words with their corresponding roots.

e. Syntax:

At this level of focus is on analyzing the words in a sentence so as to discover the grammatical structure of the sentence, revealing the structural dependency
relationships between the words. There are various grammars that can be utilized, and which will, in turn, impact the choice of a parser. All NLP applications do not require a full parse of sentences, therefore the remaining challenges in parsing of prepositional phrase attachment and conjunction scoping no longer confound those applications for which phrasal and clausal dependencies are sufficient. Syntax conveys meaning in most languages because order and dependencies are sufficient. Syntax conveys meaning in most languages because order and dependency contribute to meaning. For example the two sentences: ‘The dog chased the cat.’ and ‘The cat chased the dog.’ differ only in terms of syntax, yet convey quite different meanings. Some of the morphological ambiguities of words are resolved when the syntax of a sentence is processed and analysed well. Syntactic ambiguity is more difficult to determine.

Parsing an Arabic sentence is much more difficult than parsing an English one because of the long average length of an Arabic sentence and the complexity of Arabic syntax. Arabic sentences are more ambiguous than English sentences and there is no known or available satisfactory Arabic parser for public use. This thesis doesn’t develop one.

The process of marking words with their part-of-speech labels is known as Part-of-Speech tagging. This is easier than parsing, since it deals only with the words in the sentences and not with phrases. The tags are taken from a tag set, which is a predefined tag list.

f. Semantics:

Semantics is the study of linguistic meaning. Meaning is derived at every level. Semantic processing focuses on the interactions among word-level meanings in
the sentence and this determines the possible meanings of a sentence. This level of processing can include the semantic disambiguation of words with multiple senses. At the syntactic level, disambiguation of words that function as multiple parts-of-speech is accomplished. Semantic disambiguation permits one and only one sense of polysemous words to be selected and included in the semantic representation of the sentence. For example, amongst other meanings, ‘file’ as a noun can mean either a folder for storing papers, or a tool to shape one’s fingernails, or a line of individuals in a queue. If information from the rest of the sentence were required for the disambiguation, the semantic, not the lexical level, would be the disambiguation.

**g. Pragmatics:**

Pragmatics is the study of the use of language in situations and utilizes context over and above the contents of the text for understanding. The goal is to explain how more meanings are read into texts without actually being encoded in them. Current NLP systems put into practice modules to achieve mainly the lower levels of processing. This is for several reasons. First, there is the possibility of the applications not requiring interpretation at the higher levels. Secondly, at the lower levels there has been through research and implementation of many of the lower level concepts. Thirdly, the lower levels deal with smaller units of analysis, e.g. morphemes, words, and sentences, which are rule-governed, whereas the higher levels of language processing deal with texts and world knowledge, and which are arguably only regularity-governed and this requires much more complicated systems.
1.6 DIFFERENT APPROACHES TO NLP

There are four categories of approaches for Natural Language Processing and they are symbolic, statistical, connectionist, and hybrid.

Since the early years of development of this field, symbolic and statistical approaches have coexisted. It was in the ‘60s that connectionist NLP work initially appeared. Symbolic approaches dominated the field for a very long time. In the 1980’s, statistical approaches regained popularity as a result of the availability of critical computational resources and the need to deal with broad, real-world contexts. Following sections do some elucidation of the above-mentioned approaches.

1.6a Symbolic Approach:

This approach performs a deep analysis of linguistic phenomena and is based on unambiguous and open representation schemes and associated algorithms. The primary source of evidence in symbolic systems comes from human-developed rules and lexicons.

In logic or rule-based systems, symbolic approaches are seen. In logic based systems, the symbolic structure is usually in the form of logical propositions. Rule-based systems usually consist of a set of rules, an inference engine, and a workspace or working memory. Knowledge is represented as facts or rules in the rule-base. The inference system repeatedly selects a rule whose condition is satisfied and executed the rule.

In semantic networks fine examples of symbolic approaches can be seen. Semantic networks represent knowledge through a set of nodes that represent objects or concepts and the labeled links that represent relations between nodes and this was first proposed by Quillian. In semantic organization there is a pattern of connectivity,
that is, highly associated concepts are directly linked whereas moderately or weekly related concepts are linked through intervening concepts.

Symbolic approaches have been used for the last few decades in a number of research areas and many applications such as information extraction, text categorization, ambiguity resolution, and lexical acquisition use this approach. Typical techniques include: explanation-based leaning, rule-based learning, inductive logic programming, decision trees, etc.

1.6b Statistical Approach:

In this approach researchers employ various mathematical techniques and often use large text corpora to develop fairly accurate generalized models of linguistic phenomena based on real time examples provided by text corpora without adding major linguistic or world knowledge. In the statistical approach, researchers use observable data as the primary source of evidence.

Statistical approaches have been used basically in tasks such as speech recognition, lexical acquisition, parsing, part-of-speech tagging, collocations, statistical machine translation, and statistical grammar learning, and so on.

1.6c Connectionist Approach:

Connectionist approaches also develop generalized models from examples of linguistic phenomena which is similar to the statistical approach. The connectionist models combine statistical learning with various theories of representation and thereby the connectionist representations allow transformation, inference, and manipulation of logic formulae and this separates connectionist models from statistical models. In this system, linguistic models are harder to observe as the connectionist architectures are less constrained than statistical ones.
A connectionist model is a network of interconnected simple processing units with knowledge stored in the weight of the connections between units. Local interactions among units can result in dynamic global behavior, which leads to computation.

Some connectionist models are called localist models, assuming that each unit represents a particular concept. Localist models are quite similar to semantic networks, but the links between units are not usually labeled as they are in semantic nets. They perform well at tasks such as word-sense disambiguation, language generation, and limited inference.

Other connectionist models are called distributed models are these, unlike the localist models, have a concept called distributed models which is represented as a function of simultaneous activation of multiple units. An individual unit only participates in a concept representation. These models are well suited for natural language processing tasks such as syntactic parsing, limited domain translation tasks, and associative retrieval.

1.7 COMPARISON OF APPROACHES

From the above section, it is evident that similarities and differences exist among approaches in terms of their assumptions, philosophical foundations, and source of evidence. In addition, the similarities and differences can also be reflected in the processes, systems aspects, robustness, flexibility, and suitable tasks of each approach.

1.7a Process:

Although there are different approaches, all of them follow a general set of steps, like data collection, data analysis/model building, rule/data construction and
application of rules/data in system. The data collection stage is critical in all the approaches although statistical and connectionist approaches require much more data than symbolic approaches. In the data analysis/model building stage, symbolic approaches depends on human analysis of the data in order to form a theory while statistical approaches manually define a statistical model which is an approximate generalization of the collected data. Connectionist approaches build a connectionist model from the data. In the rule/data construction stage, manual efforts are typical for symbolic approaches. After building rules or data items, all approaches then automatically apply them to specific tasks in the system.

1.7b Systems Aspects:

In systems aspects researchers consider the source of data, theory or model formed from data analysis, rules and basis for evaluation.

Data: Symbolic approaches use human introspective data, which are usually not directly observable. Statistical and connectionist approaches are built on the basis of machine observable facts of data, usually from huge collections of texts.

Theory or model based on data analysis: A theory is formed after data analysis for symbolic approaches whereas in statistical approaches a parametric model is formed for and a connectionist model is formed for connectionist approaches.

Theory or model based on data analysis: A theory is formed after data analysis for symbolic approaches whereas in statistical approaches a parametric model is formed for and a connectionist model is formed for connectionists approaches.

Rules: In symbolic approaches, rule-construction usually results in rules with detailed criteria of rule application. For statistical approaches, rule-constructions are usually at
the surface level or under-specified. For connectionists approaches, individual rules typically cannot be recognized.

*Basis for evaluation:* Evaluation of symbolic systems is based on intuitive judgments of unaffiliated subjects. In contrast, the basis for evaluation of statistical and connectionist systems are usually in the form of scores computed from some evaluation function. However, using all will result in the task of both qualitative and quantitative evaluation.

1.7c Robustness:

Symbolic systems can anticipate anomalies by making the grammar more general to accommodate them. Statistical systems are more robust in the face of unexpected input provided that training data is sufficient. Which one needs to be sure of, but which may be difficult. Connectionist systems are also robust and fault-tolerant because knowledge in such systems is stored across the network but when presented with noisy input, they degrade gradually.

1.7d Flexibility:

Symbolic systems may lack the flexibility adapt dynamically to experience as they are guilt by human analysis of well designed examples. Statistical systems allow broad coverage compared to symbolic system, and are better at dealing with unrestricted text for more effective handling of the task at hand. Connectionists systems acquire appropriate behavior dynamically based on the give input are therefore more flexible. However, such systems find it difficult to represent structures that need to handle complex conceptual relationships, thus limiting their abilities to handle high-level NLP.
1.7e Suitable Tasks:

Symbolic approaches are suitable for phenomena that exhibit identifiable linguistic behavior. They can be used to model even at all the various linguistic levels described in earlier sections. Statistical approaches are effective in modeling language phenomena based on recurrent use of language as indicated in text corpora. Both statistical approaches and connectionist approaches can deal with linguistic phenomena that are not well understood. They are useful for low-level NLP tasks that are usually subtasks in a larger problem.

As we can see symbolic, statistical, and connectionist approaches have exhibited different characteristics and so some problems may be better tackled with one approach while other problems may be tackled better by another. In some cases, for some specific tasks, one approach may be adequate, while in other cases, the tasks can get so complex that it might not be possible to choose single best approach. As a result, researchers have begun developing hybrid techniques that make the most of the strengths of each approach in an attempt to address NLP problems more effectively and in a more flexible manner.

1.8 NLP AND THE ARABIC LANGUAGE

1.8a The Arabic Language and Arabic Dialects:

The Arabic language consists of many variations among which a particular variant enjoys a special status as the formal written standard of the media, culture and education across the Arab World. The other types of Arabic language are informal spoken dialects that are used in day to day communication. The Arabic language has many natural varieties, both historically and geographically. One group or ideology dominates the rest and so the term language is used instead of the term dialect. In the
Arab world the term Arabic language and Arab dialect is shaped by the Arab politics and religion which are nothing but primarily Arab nationalism and Islam. However there is a high degree of difference between standard Arabic and its dialects and Standard Arabic is not any Arab’s native language.

1.8b Influence of Other Languages:

In the Arab world, Algerian Arabic has a lot of influence from Berber as well as French. MS Arabic and Arabic dialects are considerably different from each other in terms of phonology, morphology, lexical choice and syntax. Factors that influence the language are geography and social class. The Arab world is divided geolinguistically in many different ways. The list below is only one of the different ways in which the Arab world is divided and that is based on the linguistic difference due to the geographic placement of the place.

- Egyptian Arabic includes the dialects of the Nile valley: Egypt and Sudan.
- Levantine Arabic is the dialects of Lebanon, Syria, Jordan, and occupied Palestine.
- North African Arabic called Maghreb Arabic covers the dialects of Morocco, Algeria, Tunisia and Mauritania. Libyan Arabic is sometimes included.
- Gulf Arabic is spoken in Kuwait, United Arab Emirates, Bahrain and Qatar. Saudi Arabia is usually included although there is a wide range of sub-dialects within it. Omani Arabic is included some times.
- Iraqi Arabic has elements of both Levantine and Gulf.
- Yemenite Arabic is a separate class of its own.
• Maltese Arabic is not always considered an Arabic dialect. It is the only Arabic variant that is considered a separate language and is written with the Roman script.

1.9 CHARACTERISTICS OF ARABIC

It is common to distinguish three sub-dialects within each dialect region: city, rural and Bedouin which is socially influenced. The three degrees are often associated with a class hierarchy in which rich settled city dwellers are on top and Bedouins are of the lower class. Just as in other languages in Arabic also different social associations exist. For example, the city dialect is considered less marked and more refined and prestigious; whereas the Bedouin dialect is considered less prestigious and crude, yet the origin is considered pure. Arabic is spoken with alternate variants by the speakers in different social contexts.

There is a special kind of coexistence between MSA and the dialect specific to the region though the relationship between them is considered complex. This kind of situation is what linguists call diglossia. Though both the variants have clear domains of prevalence yet there is a large gray area in between that is often filled with a mix of the two forms.

Arabic language has two major characteristics that make it a very complicated language and they are the agglutinative nature of the language and the vowellessness nature of the language which create a log of ambiguity while translating. So the language has to be pre-treated in order to disambiguate the contents. In the following section we will study the aforementioned natures so that a clearer picture emerges.
1.9a Agglutinative Nature:

Basically in the Arabic language it is very difficult to identify words in sentences. Most European languages consider word as a string of characters between two blanks. So for the machine to identify the word the blanks serve as markers. What appears as a word in other languages happens to be affixes in Arabic and these affixes can be part of the word after the blank. This makes the words all the more ambiguous as there could be a multiword with one affix and as it looks like two or three separate words in English whereas the same can be written in Arabic as follows:

The following Table 1.2 shows some examples of agglutinative Arabic words.

Table 1.2 Arabic words that are Agglutinative in nature

<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>liyaktubuu</td>
<td>Let them write</td>
</tr>
<tr>
<td>yuqatiluuhum</td>
<td>They are fighting them</td>
</tr>
<tr>
<td>yunaSruhum</td>
<td>They are supporting them</td>
</tr>
</tbody>
</table>

1.9b Prefixes:

A number of prefixes in Arabic can be placed in the following grammatical categories in English.

Table 1.3

<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>al</td>
<td>the (article)</td>
</tr>
<tr>
<td>fa, waw and &amp;um</td>
<td>and (connectives)</td>
</tr>
<tr>
<td>bi</td>
<td>with (preposition)</td>
</tr>
<tr>
<td>li</td>
<td>for (preposition)</td>
</tr>
<tr>
<td>ka</td>
<td>to (preposition)</td>
</tr>
<tr>
<td>li</td>
<td>(imperative)</td>
</tr>
</tbody>
</table>
Although the particles here are with vowels, in the written language the vowels are not always included. In Arabic there are words with three consonants but again this is not always the case.

There are some morphological constraints on using these prefixes. For example, *sa* is the practice of future and it cannot be used with a noun or an adverb or any other grammatical category. So these kinds of constraints bring down the range of words that can be prefixed with this particle. In Arabic many words which are nouns can resemble verbs if they are not vocalized. So for the computer to identify the prefixed word a different marking is used. This enables the software to give a translation of a higher quality. There is a need to build a huge database in order to determine the prefixes in Arabic.

**1.9c Suffixes:**

There are seventeen possessive suffixes in Arabic apart from the undefined accusative suffix and the suffix of the energetic. Possessive suffixes consist of one or two consonants. There are combinations of suffixes that remain ambiguous. For example, the suffix *ha* making third person feminine can be confused with the undefined accusative of a word that ends with *h*.  

Accusative NPs with adjectives in them end with ‘alif” and adverbs also end with ‘alif”. Unless clear distinctions are made between these two categories, it is very difficult to process the language during translation. Though it is possible to make these distinctions in the lexical database, it is only the context that can serve as an effective measure in the corpora and so the difficulty level arises at this state.
Examples of suffixes Table 1.4.

### Table 1.4 Suffixes

<table>
<thead>
<tr>
<th>Suffix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ii</td>
<td>addressee: second person singular feminine</td>
</tr>
<tr>
<td>aa</td>
<td>addressee: second person dual mas/fem</td>
</tr>
<tr>
<td>uu</td>
<td>addressee: second person plural mas</td>
</tr>
<tr>
<td>na</td>
<td>addressee: second person plural fem</td>
</tr>
</tbody>
</table>

Examples of adjectives and adverbs that end in “alif” as follow in Table 1.5.

### Table 1.5 Sample sentences with adverb and adjectives

<table>
<thead>
<tr>
<th>Word</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>kitabhA</td>
<td>her book</td>
</tr>
<tr>
<td>makanhA</td>
<td>at her place</td>
</tr>
</tbody>
</table>

1.9d The Vowelless Nature of Arabic:

In Arabic corpora the vowelles nature of words causes problems at the word level and the sentence level. This causes problems in the derivation of word meaning and the grammar especially in the case of verbs. Let us take the word *d-r-s* ‘notion of study’. It has four possibilities as given in Table 1.6.

### Table 1.6

<table>
<thead>
<tr>
<th>Word</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>darasat</td>
<td>first person singular</td>
</tr>
<tr>
<td>darasat</td>
<td>second person singular masculine</td>
</tr>
<tr>
<td>darasat</td>
<td>second person singular feminine</td>
</tr>
<tr>
<td>darasat</td>
<td>third person singular feminine</td>
</tr>
</tbody>
</table>

Now in this case it is next to impossible to have a computer decide the actual grammar as this can be decided only if we have someone understand the context and then deal with the grammar based on the context.
Some examples are given where a combination of two consonants can have more than one meaning: Table 1.7.

**Table 1.7 Combination of two consonants**

<table>
<thead>
<tr>
<th>Consonants</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>laka</code></td>
<td>for you</td>
</tr>
<tr>
<td><code>laka</code></td>
<td>hit with the fist</td>
</tr>
<tr>
<td><code>kl</code></td>
<td>all</td>
</tr>
<tr>
<td><code>kl</code></td>
<td>to become tired</td>
</tr>
<tr>
<td><code>haDa</code></td>
<td>this</td>
</tr>
<tr>
<td><code>haDa</code></td>
<td>to talk irrationally</td>
</tr>
</tbody>
</table>

Statistics prove that in 90% of the times for the above examples the first meaning is taken into consideration as this is the meaning that is more frequently used. So if we use statistics in the translation software the probability of getting the correct meaning is higher than using the second meaning and this helps in understanding the importance of tagging the corpora together with text which is very valuable in determining the different categories of forms and words.

Both agglutination and vowellessness of the Arabic language go together in such a way that their combination complicates the identification of words. Badawi, Carter and Gully (2004) afford a glimpse of various possible interpretations for one set of characters between blanks.

### 1.10 TOKENIZATION IN ARABIC

Tokenization is a very important problem as it is “closely related to morphological analysis” (Chanod and Tapanainen 1996). In a language like Arabic which has a complex morphology, it is even more important. The function of a tokenizer is to split a running text into tokens, so that they can be fed into a
morphological transducer or POS tagger for further processing. It is therefore understood that a tokenizer is responsible for defining word boundaries, demarcating clitics, multiword expressions, abbreviations and numbers.

Clitics are bound morphemes that have syntactically relevant distribution. Typically, affixes carry morpho-syntactic features (such as tense, person, gender or number) while clitics serve syntactic functions (such as negation, definition, conjunction or preposition) that would otherwise be performed by an independent lexical item. Tokenization is thus a very important step for a syntactic parser that needs to build a tree from syntactic units. An example of clitics in English is the genitive suffix ‘s’ in the teacher’s bag.

Arabic clitics are not easily recognizable. Clitics use the same alphabet as words with no demarcating mark as the English apostrophe, and they can be concatenated one after the other. To detect and mark clitics one needs to have morphological knowledge. Arabic Tokenization has been described in various researches and implanted in many solutions as it is a required preliminary stage for further processing. None of the projects, however, shows how multiword expressions are treated, or how ambiguity is resolved.

1.10a Arabic Tokens:

A token can be word, a part of a word (or a clitic), a multiword expression, or a punctuation mark in any language especially in Arabic. A tokenizer needs to know word boundaries, such as white spaces and punctuation marks, and also information about the token boundaries inside words when a word is made up of a stem and clitics. All main tokens are delimited either by a white space or a punctuation mark.
Full form words can then be divided into sub-tokens, where cities and stems are separated.

1.10b Main Tokens:

A tokenizer relies mainly on white spaces and punctuation makes as delimiters of word boundaries (or main tokens). Additional punctuation marks are used in Arabic such as the comma, question mark and semicolon. Numbers are also considered as main tokens. While most Arab countries use the Hindi numerals, a few Arab countries use the Arabic numerals as in English. Therefore a list of all punctuation marks and number characters must be fed to the system to allow it to demarcate main tokens in the text.

1.10c Sub-Tokens:

Arabic morphotactics allows words to be prefixed or suffixed with clitics (Attia 2006). Clitics can be concatenated one after the other. Also, clitics undergo assimilation with word stems and with each other, which makes them even more difficult to handle in any superficial way. A verb can comprise up to four sub-tokens (a conjunction, a complementizer, a verb stem and an object pronoun) as show in Figure 1.1.

![Figure 1.1 Verb](image-url)
Similarly a noun can comprise sub-tokens, usually up to four though the Figure 1.2 shows five sub-tokens. The define article and the genitive pronoun are mutually exclusive sub-tokens.

Figure 1.2 Noun

Apart from what we have seen above one should understand that there are numerous rules that govern the combination of words with affixes and clitics, which are called grammar-lexis specifications (Dichy and Fargaly 2003). The rule that states that adjective and proper nouns do not combine with possessive pronouns is an example at the grammar-lexis interface.

Many natural languages including Semitic languages use finite state technology which has been very successfully used in developing morphologies and text processing tools. Lexical entries along with all possible affixes and clitics are encoded in the lexc language which is a right recursive phrase structure grammar (Beesley and Karttunen 2003) in a standard finite system. A lexc file is that which contains numerous lexicons connected through what is known as “continuation classes” which determine the path concatenation. According to Mohammed A. Attia there are a number of advantages in this technology that make it especially attractive in dealing with human language morphologies.
Among these advantages are:

- The technology is fast and efficient. It can handle very huge automata of lexicons with their inflections. Compiling large networks that include several millions of paths is only a matter of second in a finite state calculus. Moreover, these large networks can be easily combined together to give even larger networks.
- Handling concatenative and non-concatenative morphotactics (Beesley 1998)
- Unicode support, which enables developers to accommodate native scripts that use non-Latin alphabets.
- Multi-platform support. Xerox finite state tools work under Windows, Linux, UNIX and Mac OS, which means that a morphological transducer developed using Xerox finite state compilers can serve applications under any of these platforms.
- A finite state system is fully reversible. So it can be used for analysis as well as generation.
- The regular expressions used in finite state systems closely resemble standard linguistic notations (Yona and Wintner 2007). The rules are thus reasonably readable and intelligible.

1.11 TOKENIZATION SOLUTIONS

Depending on the depth of the linguistic analysis involved, there are different levels at which an Arabic tokenizer can be developed. The tokenizer relies on white spaces and punctuation marks to demarcate main tokens. In demarcating sub-tokens, however, the tokenizer needs more morphological information. This information is provided either deterministically by a morphological transducer, or indeterministically
by a token guesser. Eventually both main tokens and sub-tokens are marked by the same token boundary, which is the same sign everywhere in the input given to the software. The classification into main and sub-tokens is a conceptual idea that helps in assigning the task of identification of different components.

Tokenization is a process that is connected to and dependent on morphological analysis. There are different models in the market which are dependent on the levels of linguistic depth. The tokenizer also interacts with other components helping to resolve the complexity and ambiguity issues. Tokenization output is possible only if the token filters are applied.

1.12 THE CHALLENGES FACED BY RESEARCHERS IN ANLP (ARABIC NLP)

The Arabic language is both challenging and interesting because of its complex linguistic structure (Attia 2008), due to the history and importance of its people, the region they occupy, and its cultural and literacy heritage (Bakalla 2002).

In the last few years, Arabic natural language processing (ANLP) has gained a lot of importance, and for a wide range of applications, including machine translation, information retrieval and extraction, speech synthesis and recognition, localization and multilingual information retrieval systems, text-to-speech, and tutoring systems. Many state-of-the-art systems have been created. All these applications had to deal with several complex problems pertinent to the nature and structure of the Arabic language as we have seen earlier in this thesis. Many of the ANLP systems are developed keeping in mind the Western world focus on tools that enable non-Arabic speakers to understand Arabic texts. Arabic Tools such as machine translation and sentiment-analysis are very useful to intelligence and security agencies and due to the
great demand for such tools that existed, these tools were developed using machine learning approaches. This approach is fast and inexpensive. There were a few difficult issues for the developers to deal with and one such problem is the inclusion of Arabic texts which were either translated and transliterated having inconsistent spellings (Shaalan and Raza 2008).

Example:

<table>
<thead>
<tr>
<th>Arabic Equivalent</th>
<th>English name</th>
</tr>
</thead>
<tbody>
<tr>
<td>washintun</td>
<td>Washington</td>
</tr>
</tbody>
</table>

Some of the challenges that researchers face in ANLP are given below:

A. Significance of ANLP in the West and in Arab Countries:

Funding for the development of ANLP applications has surged in the U.S. since September 11, 2001. The U.S. Department of Homeland Security was confronted with very difficult tasks ranging from identifying Arabic names correctly at airport security and in Arabic documents seized by the American authorities in the U.S. and abroad. They had also accumulated an enormous volume of Arabic texts that they had no clue as to whether they were relevant or not. They had neither the human expertise needed to perform the task nor the time to wait for human translators to complete the task. ANLP tools that could scan such documents to recognize names, places, dates, etc., of interest soon became essential. As a result, funding became available for companies and research centers to develop tools such as named entity recognition, machine translation, especially spoken machine translation, document categorization, etc.
B. Objectives of ANLP:

ANLP applications developed in the Arab World have different objectives and usually employ both rule-based and machine-learning approaches. Some of the objectives ANLP for the Arab World are given below:

a. *Transferring knowledge and technology from the West to the Arab World:* Most recent publications in science and technology that are published in the English language are not available to Arab readers who have very little or sometimes no competence in English. To use human translators to translate such an enormous amount of data to Arabic is very costly and time consuming. So Arabic NLP could help reduce the time and cost of translating, summarizing, and retrieving information in Arabic for Arab speakers.

b. *Modernising the Arabic language:* Translating new concepts and terminology into Arabic involves coinage, arabization, and making use of lexical gaps in the Arabic language. This will definitely affect that revitalization of the Arabic language and enable it to fulfill the essential needs for its speakers.

c. *Modernising Arabic Linguistics:* Arabic NLP needs a more formal and precise grammar of Arabic than the complex traditional grammar so widely employed today. One needs to create a new system which helps in preserving the valuable heritage of traditional Arab grammarians. Our treatment of Arabic imperatives could be a case in point.

d. *Making Information Recovery, Extraction and Translation Available for the Arabs:* Bridging the gap between peoples of the Arab world and their peers in more technically advanced countries is the basic aim. If information is available to Arabic speakers in their native language, Arabic NLP tools gives power to the present generation of erudite Arabs. In order to attain parity with
the rest of the world, Arabic NLP tools are indispensable, which is a matter of national security to the Arab World (Farghaly 2010)

C. Challenges:

1. The Diglossic Nature of Arabic:

The Arabic language has many variants and this phenomenon where there are more than two variations in the speech of the same language is called ‘diglossia’. Usually most languages have one formal way of speech and many informal ways of speech, which are region-specific or community-specific.

When do we call a language ‘diglossic’ in nature? First of all if a language uses a high variety of speech as in official speech, news broadcasts etc, as opposed to a low variety, then this language can be termed ‘diglossic’. There are a number of languages that fall into this classification including Arabic. Secondly a speaker of such high variety of language is considered highly educated and qualified. This classical variety of Arabic is associated with religion and it is part of Quran and so it is prestigious for the speaker to be considered as a person with such high knowledge. Thirdly a lot of literature is written in this type of language in any diglossic language. Fourthly the way in which people acquire each variety of the language is very important to the language itself. At home children speak the lower level “naturally” but in the discourse or formal speech situation, it’s the higher level spoken by the speaker which is acquired after a number of years of learning and practicing the language.

For over 1500 years this diglossic nature of the language has existed in Arabic. But this can’t be considered as the standardization process of the Arabic language. According to Ferguson (1959) diaglossia is:
“a relatively stable language situation in which, in addition to the primary dialects of the language, (which may include a standard or regional standards), there is a very divergent, highly codified (often more grammatically complex) superposed variety, the vehicle of large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community ordinary conversation.”

There are of course, significant implications for developing NLP systems in a diglossic situation like Arabic. First, it is very difficult and almost impossible for any one ANLP application to process data from all the varieties of Arabic. Each variety has its own grammar, lexicon, and morphology even though they have some properties in common. An ANLP application has to specify beforehand which variety it is aiming to address. Moreover, the application has to have a good “understanding” of the linguistic properties of the particular variety it aims at. An understanding of the complex sociolinguistic situation of Arabic can be very useful for ANLP researchers and developers.

Research in developing NLP tools have focused on handling texts that are primarily written in MSA. However, when these tools are applied on the classic text or spoken language from various regions, accuracy is not there in grammar, syntax, and expressions from one variety to another and between the dialects themselves. It is also to be noted that there are few resources available in the form of grammars or dictionaries as the basis for developing NLP for the dialects.
Researchers at Columbia University (Habash et al. 2005) in their approach to overcome this problem have assumed that it is simpler to develop NLP systems for the dialects by first extracting and categorizing the systematic grammatical features of a dialect, making it more like MSA and then applying MSA natural language processing tools to process a text. Another approach is to create Dialect Treebanks that resemble MSA Treebanks by exploiting systematic regularities within a dialect and among dialects. For example the work reported in Shaalan et al (2007) transfers Egyptian Arabic texts to MSA using a lexical transfer approach in addition to changing the SVO Egyptian order into the MSA VSO order. They also enhanced the tables of Buckwalter’s Morphological Analyzer to transform Egyptian Arabic words into MSA words. Following the same approach one could also reuse MSA tools to process Colloquial Arabic.

2. The Script:

The Arabic script itself is a big challenge for converting Arabic into automatic processing due to its complex phoneme-grapheme relations. Arabic has the advantages of being a language that has a specific letter for each sound and can be considered as a phonetic language. But if, however, we take short vowels, for example, there is no written representation or script for it. So one cannot consider this as an easy language and sometimes the form of the letter changes depending on where it is placed in the word. The other problem faced by researchers is that in this language, there is not capitalization and this is one language that has minimum punctuation.

As an example of the regularity of the association of letters with sounds, the letter $k$ in Arabic is always pronounced as $ka$, unlike letters in English that have more than on pronunciation. For example, the letter ‘c’ in English may be pronounced as /s/
or /k/ as in ‘cycle’. Also English has silent letters such as the ‘k’ in ‘knife’ the ‘b’ in ‘thumb’, and the ‘g’ in ‘gnat’ while Arabic has no silent letters. In Arabic, a combination of two letters does not produce a new sound. For example, combining the letters ‘t’ and ‘h’ in English sometimes produces a voiceless interdental fricative as in ‘though’ or a voiced interdental fricative as in ‘though’.

In Arabic there are no dedicated letters for short vowels: short vowels are represented by diacritics which are marks above or below the letters. In contemporary writings these diacritics are not used much and the readers are expected to fill in the missing short vowels through their knowledge of the language. It is difficult for non native speakers to read the text and this poses a great challenge for automatic processing of Arabic.

We have earlier seen that Arabic letters have different shapes depending on the position of the letter in the word. In the following table we can see the shape of the initial letter, median letter, final letter and the non connecting letter.

<table>
<thead>
<tr>
<th></th>
<th>ﺧ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>ﺧ</td>
</tr>
<tr>
<td>Median</td>
<td>ﺧ</td>
</tr>
<tr>
<td>Final</td>
<td>ﺧ</td>
</tr>
<tr>
<td>Non-connecting</td>
<td>ﺧ</td>
</tr>
</tbody>
</table>

There are rules governing the selection of the shape of the letter which is related to the position it occupies in the word. Arabic word processors put into practice these rules so that the user can depend upon automatic selection of the correct shape. Therefore there is only one key for each letter and there are rules encoded to recognize the context and insert the correct shape automatically. Moreover, there are shapes which the morphological processing tool should handle.
Non-Latin script do not follow the capitalization and usage of uppercase and lowercase script, but in English and other Latin script-based languages, most sentences begin with an uppercase letter and end with a period. In the applications used for machine translation, information retrieval, clustering, and classification, it is necessary to split a running text correctly into sentences and a sentence-splitter capitalizes on these features as these soft-wares are programmed thus. This is definitely a problem for scripts such as Arabic, Chinese, Japanese and may other non-Latin based script that have neither capitalization nor strict rules of punctuation. It is very difficult to recognize sentence boundaries in languages like Arabic, and sometimes in Arabic an entire discourse can be written with only a period at the end of the paragraph.

Arabic discourse is characterized by excessive use of coordination, subordination and logical connectives. Some of the Arabic coordinators that conjoin a sentence are given below:

<table>
<thead>
<tr>
<th>Arabic</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>wa</td>
<td>and</td>
</tr>
<tr>
<td>fa</td>
<td>and</td>
</tr>
<tr>
<td>&amp;um</td>
<td>then, and</td>
</tr>
</tbody>
</table>

Capitalization and sentence boundaries help in recognizing names also. This is called named Entity Resolution (NER) (Shaalan and Raza 2009). In the absence of capital letters for proper nouns, it is necessary to evolve means to recognize proper nouns. While extracting information the system should be able to extract unstructured information like postal address, titles of people, cities etc.
Standardisation of the Arabic Script:

Researchers and software developers face a challenge trying to standardize the Arabic script. The inconsistency in the usage of diacritics and certain letters raises problems in establishing a standard. There are Arabic letters that have same shapes and a dot above or below differentiates the letters.

Though there has been standardization of Arabic with MS Arabic, texts written in MSA do not include vowelling and the marks above or below the script is not available either. So ANLP manages this by standardizing the input text as given below:

<table>
<thead>
<tr>
<th>alif madda</th>
<th>bare alif</th>
</tr>
</thead>
<tbody>
<tr>
<td>alif hamza</td>
<td>bare alif</td>
</tr>
<tr>
<td>taa marbuu ta</td>
<td>final haa</td>
</tr>
<tr>
<td>alif maqsuura</td>
<td>yaa</td>
</tr>
</tbody>
</table>

To handle this situation, the common practice in Arabic NLP systems is to normalize the input text (Larkey and Connell 2001). For example, in order to handle the different variations in Arabic script, Larkey and Connell (2001) replace the intial alif with a hamza above or below with simply an alif, or bare alif. They also normalize the alif madda with a bare alif. Further, they normalize the final taa marbuuTa with a final haa and the alif maqsuura with the yaa.

The Stanford Arabic Statistical Parser has a similar approach. It was designed by the Stanford Natural Languages Processing Group. Farghaly and Senellart in the SYSTRAN Arabic-to-English machine translation system, also incorporated normalization. But it soon became apparent that although normalization improves recognition by solving the variability in input, it increases the probability of ambiguity (Farghaly 2010).
Ambiguity in ANLP:

Though there have been several attempts at normalization of the script, this has only increased the levels of ambiguity. There are many levels of ambiguity and this poses a greater challenge to researchers and developers (Attia 2008). As we have seen in the last section, ambiguity exists at various levels and according to SYSTRAN the average number of ambiguities for a token is as high as 19.2 in MSA, while in other languages it is only 2.3. Even though due to the absence of short vowels in MSA, there is ambiguity is also because of the complexity of the language itself.

Ambiguities in the language are at the following levels:

- homograph
- internal word structure
- syntactics
- semantics
- constituent boundary
- Anaphora

Explanation of the above:

1. **Homographs**: A word belonging to more than one part of speech such as *qdm* which could be a verb of form II meaning ‘to introduce’ or a verb of form 1 meaning ‘to arrive from’ or a noun meaning ‘foot’. Some homograph ambiguity can be resolved by contextual rules. For example, an Arabic word that could be either a noun or a verb can be disambiguated by the following rule which says that such a word will be disambiguated to a noun when preceded by a preposition.
(2) **Internal word-structure ambiguity:** That is, when a complex Arabic word could be segmented in different ways. In English, hammer could be one monolithic word, meaning ‘an instrument’ or it could be ham-er, ham meaning ‘to overact’.

(3) **Syntactic ambiguity:** As in the case of a prepositional attachment is in *qabaltu mudiyr al-bank al-jadiyd* which could mean ‘I met with the new bank manager’ or ‘I met with the manager of the new bank’ depending on the internal analysis of the noun phrase.

(4) **Semantic ambiguity:** Sentences and phrases may be interpreted in different ways. For example, *yuḥb khalid ahmad ak&r mn ibraham* ‘Khalid likes Ahmed more than Ibrahim.’ Does this mean that Khalid likes Ahmed more than Ibrahim, or do Khalid and Ibrahim like Ahmed, but Khalid likes Ahmed more than Ibrahim likes Ahmed?

(5) **Constituent boundary ambiguity:** For example, *mudyr al-bank al-jadiyd* could mean ‘the new manager of the bank’ or ‘the manager of the new bank’ depending on the boundary of the adjective phrase within this noun construct.

(6) **Anaphoric ambiguity:** As in *qAla Ali annahu najah* ‘Ali said that he succeeded’. This sentence is ambiguous both in English and Arabic. Chomsky’s Binding principles account for sentences like this. The question here is does ‘he’ refer to Ali or to someone else? Another interesting example is the following:

The journalist, met the minister, who criticized him.

(Who criticized who?)

There are two lexical NPs: the journalist and the minister. Each has a different index. However, each of ‘who’ and ‘him’ can refer to any of the two lexical NPs. This is represented by the indices each has. Both have the indices $i$ and $j$ indicating the ambiguity they have in that each may refer to the journalist or to the minister.
Only discourse information could disambiguate such sentences. In addition to these levels of ambiguity the process of normalization plus features of Arabic such as the pro-drop structure, complex word structure, lack of capitalization, and minimal punctuation contribute to ambiguity, but it is the absence of short vowels that contributes most significantly to ambiguity. With the absence of short vowels, two types of linguistic information are lost. The first is most of the case markers that define the grammatical function of Arabic nouns and adjectives. For example, a Damma, which is a high back rounded vowel at the end of a common noun or adjective marks the nominative case whereas a fatHa, which is a low front vowel in the final position of a common noun, marks the accusative case and a kasra which is a high front vowel marks the genitive case. The absence of case markers and thus the grammatical function of a word, creates multiple ambiguities due to the relatively free word order in Arabic and because Arabic is a pro-drop language.

Ambiguity also exists in the lexical and part-of-speech information in the Arabic language and this is due to the absence of internal vowelling. Without contextual clues it is difficult to determine the part-of-speech of a word in the Arabic language. Contextual clues could be a preposition meaning, a wh-phrase meaning or a verb meaning. While ambiguity is a challenge in any language, it makes for a bigger change in Arabic due to all these features present in the language.

Tokenization works very well for non-agglutinative languages like English where a token consists of a sequence of one or more letters preceded and followed by space. A single Arabic word may consist of four different tokens and so tokenization requires prior knowledge of the constraints on concatenating affixes and clitics within Arabic words. There is a definite distinction between clitics which are syntactic units and thus have their own part of speech but do not stand alone, and affixes that mark
grammatical inflections such as tense, number and person agreement. If this knowledge is there, then the process becomes a little easy for developers and researchers.

The Syntactic Structure of Arabic:

Word order in Arabic is relatively free compared to many other languages. Though both Classical Arabic and Modern Standard Arabic have a word order that is primary verb-subject-object (VSO), they also allow subject-verb-object (SVO) and object-verb-subject (OVS). In headlines SVO is a common feature. Arabic dialects exhibit SVO order. Arabic has a very rich and complex agreement system. A noun and its modifiers have to agree in number, gender, number and person. However, in VSO sentences the verb is always in the singular even when its subject is dual or plural. The feature ‘definiteness’ plays an important role in constituent formation.

SVO structure

Indefinite noun followed by

a definite noun.

VSO structure

Classical Arabic grammar was developed between the 8th and 10th centuries. This was developed to keep the purity of the Arabic language intact as there was a fear of influence from other languages due to the Islamic conquests that took place back then. Due to the absence of Classical Arabic reference grammar and lexicon, it is difficult to determine which usage is correct. The evidence of short vowels and explicit case marking is very much present in Quran and so Arab grammarians had to give explanations for case markings in their grammatical texts.
The texts seen in today’s media like newspaper, journals and books the usage of MSA is evident and these texts distinctly display the absence of short vowels and case makings.

**D. Solutions:**

Robust, and not so robust, grammatical descriptions of Modern Standard Arabic have begun to appear (Badawi et al. 2004). Such descriptions are very useful in the processing of contemporary Arabic although they were not written from a computational viewpoint. Annotated Arabic corpora that have been developed at the LDC (Linguistic Data Consortium, Stanford) are extremely valuable for ANLP applications. Modern Standard Arabic texts have been analyzed with insights from traditional Arabic grammar as well as from modern linguistic theories. The LDC has also compiled corpora for some Arabic dialects and Arabic-English parallel corpora that are very useful for machine translation. Recently, as we mentioned earlier, the LDC released an annotated entity extraction corpus for Arabic. Another important resource is the Prague Arabic Dependency Treebank which implements a functional approach to the analysis of Modern Standard Arabic. There are also resources for Arabic dialects such as the Arabic Treebank at Columbia University and the Arabic dialects corpora at the LDC.

**1.13 SUMMARY**

In this chapter we have elucidated the nature of NLP, its various levels and taxonomies. This is important as this helps a researcher to understand NLP in its totally and will take a researcher to greater heights of research leading him to discover newer models that are faster and efficient using technology to its advantages.