AL-AZHAR UNIVERSITY
FACULTY OF LANGUAGES AND TRANSLATION
ENGLISH DEPARTMENT

Implications of the Agreement Features in Machine Translation

An M.A. Thesis

Submitted by

Mohammed Abdel-Aal Attia

Under the Supervision of

Dr. Muhammad Mahmoud Ghali,
Professor of Linguistics,
Faculty of Languages and Translation,
Al-Azhar University

Dr. Ahmed Shafik Elkhatib,
Professor of Linguistics,
Faculty of Languages and Translation,
Al-Azhar University

Dr. Muhammad Kamal El-Din Abdel-Ghani,
Assistant Professor of Linguistics,
Faculty of Languages and Translation,
Al-Azhar University

1423 A.H./2002 A.C.

Acknowledgment

I would like to extend my thanks and appreciation to all those who helped me in completing my research.

My appreciation is due, of course, to my supervisors: Dr. Muhammad Mahamoud Ghali for his continual assistance and support, Dr. Ahmed Shafik Elkhatib for his insightful remarks and meticulous revision of the first draft of this thesis, and Dr. Muhammad Kamal El-Din Abdel-Ghani for nurturing my interest in studying the phenomenon of agreement in English and Arabic in a comparative way.

My deep appreciation is due to Harf Information Technology and Sakhr Software companies for shaping and widening my knowledge of computers, programming, and translation technologies during the period I worked for them.

My thanks are also due to Mr. Shourbaji Mohammed Shourbaji, a specialist in Arabic, for his great help in revising the parts related to Arabic grammar.

I would also like to thank the people in the Faculty of Languages and Translation Library and Al-Azhar Central Library for keeping their libraries up to date in the field of Linguistics.

I would also like to thank the people in the AUC Library for giving me access to their invaluable books and facilities.

Last but not least, I would like to extend my appreciation to my wife for her continuous encouragement and for allowing me the time and providing the supportive environment needed for study, and to my little daughters, Menna and Nada, for the inspiration and motivation they granted to me.

Table of Contents

List of Figures

List of Tables

Transliteration of Arabic Sounds[1]

Letter (E)	Transliteration	Letter (A)	Phonetic Description[2]
hamzah	ʾ	أ	voiceless glottal stop
bāʾ	b	ب	voiced bilabial stop
tāʾ	t	ت	voiceless apico-dental stop
ṯāʾ	ṯ	ث	voiceless inter-dental fricative
ǧīm	ǧ	ج	voiced lamino-alveolar palatal affricate
ḥaʾ	ḥ	ح	voiceless radico-pharyngeal fricative
ḫaʾ	ḫ	خ	voiceless dorso-uvular fricative
dāl	d	د	voiced apico-dental stop
ḏāl	ḏ	ذ	voiced inter-dental fricative
rāʾ	r	ر	voiced apical trill roll
zāy	z	ز	voiced apico-alveolar fricative
sīn	s	س	voiceless apico-alveolar fricative
šīn	š	ش	voiced lamino-palatal fricative
ṣād	ṣ	ص	voiceless apico-alveolar emphatic fricative
ḍād	ḍ	ض	voiced apico-dental emphatic fricative
ṭāʾ	ṭ	ط	voiceless apico-dental emphatic stop
ẓāʾ	ẓ	ظ	voiced inter-dental emphatic fricative
ʿaīn	ʿ	ع	voiced radico-pharyngeal fricative
ġaīn	ġ	غ	voiced dorso-uvular fricative
fāʾ	f	ف	voiceless labio-dental fricative
qāf	q	ق	voiced dorso uvular stop
kāf	k	ك	voiceless velar stop
lām	l	ل	voiced apico-alveolar lateral
mīn	m	م	voiced bilabial nasal
nūn	n	ن	voiced apico-alveolar nasal
hāʾ	h	هـ	voiced laryngeal fricative
wāw	w	و	voiced bilabial (rounded) velar glide
yāʾ	y	ي	voiced palatal (unrounded) glide

Short Vowels

fatḥah	a	ـَ
kasrah	i	ـِ
ḍammah	u	ـُ

Long Vowels

ā
ī
ū

Compound Vowels

aw
ai

[1] We follow the DIN 31635 standard for the transliteration of the Arabic alphabet.
[2] Usama Mohamed Soltan, "A Contrastive and Comparative Syntactic Analysis of Deletion Phenomena in English and Standard Arabic", unpublished dissertation, Ain Shams University, 1996, p. vii.

List of Abbreviations and Acronyms

Abbreviation	Full Form
A	Arabic
ACC	Accusative (case)
adj	adjective
adv	adverb
ALPAC	Automatic Language Processing Advisory Committee
AP	Adjectival Phrase
ASCII	American Standard Code for Information Interchange
aux	auxiliary
C	Complex (transitive verb)
cat	(grammatical) category
compl	complement
conj	conjunction
D	Ditransitive (verb)
DB	Database
dec	declarative
det	determiner
dl	dual
E	English
EBMT	Example-Based Machine Translation
f	female (or feminine)
FAHQMT	Fully Automated, High Quality Machine Translation
fn	finite (sentence)
fun	function
fut	future (tense)
GEN	Genitive (case)
gend	gender
I	Intransitive (verb)
IBM	International Business Machines Corporation
inan	inanimate
ind	indicative
int	interrogative
IP	Inflectional Phrase
ISO	International Organization for Standardization
KBMT	Knowledge-Based Machine Translation
L	Linking (verb)
lex	lexical item
m	male (or masculine)
md	mood
MIT	Massachusetts Institute of Technology
mod	modifier
MT	Machine Translation
n	noun

NLP	Natural Language Processing
NOM	Nominative (case)
NP	Noun Phrase
nt	neuter
num	number
obj	object
OVS	Object-Verb-Subject
pas	past (tense)
PC	Personal Computer
pers	person
pl	plural
POS	Part of speech
pospron	possessive pronoun
PP	Prepositional Phrase
pred	predicate
prep	preposition
pres	present (tense)
pron	pronoun
qtf	quantifier
SA	Standard Arabic
sem	semantic (feature)
sg	singular
SL	Source Language
Spec	Specifier
subcat	(grammatical) subcategory
subj	subject
SVO	Subject-Verb-Object
T	Transitive (verb)
TL	Target Language
tns	tense
v	verb
VOS	Verb-Object-Subject
VP	Verb Phrase
VSO	Verb-Subject-Object

Introduction

In recent years, research as well as software applications have been concentrating on Machine Translation (MT). This is due to many factors, the most important of which is the increasing need to create online communication between different parts of the world and between people speaking different languages. People felt the need for machine translation since the advent of computers, but the early attempts they made were completely dissatisfactory. It was based on a primitive idea of processing the source text through an electronic dictionary that included words of the source language and their equivalents in the target language, with no further manipulation either of the input or the output. The result they received at that time was disappointing. This led research in this field to be blocked for nearly a decade. However, with the modern insights in Linguistics and software engineering, natural language processing systems have witnessed remarkable advances.

Translating a human language to another one through the computer is never an easy task. A human language is a highly complicated system, and so MT involves a big deal of complicated manipulation and analysis. Despite the great advances done in the field of Computational Linguistics, MT is not accomplished and is still far from being satisfactorily accomplished.

In the MT process, first, words of the input text are identified and analyzed morphologically to know where each word belongs. We need to know that *mice* is the plural form of *mouse* and that *went* is the past of *go*. Then, based on the morphological knowledge, parsing of the lexical structures takes place. Now the Noun Phrase, Verb Phrase, and other phrases of the sentence are identified. Where one phrase ends and the other begins is clearly defined. Grammatical functions (who is doing what to whom) are also determined. Then the sense that most suits the word in this context is selected from among a large variety of choices. At this stage, sense disambiguation and idiom and phrasal verb detection are all manipulated.

Once the source language is completely analyzed and understood, the generation into the target language starts. Representations at the lexical, syntactic, and semantic levels are transferred into the target language. After meaning is transferred, a good deal of manipulation in the target language is required. Words are to be placed in their correct order, the order that most suits the target language. After words are placed in order,

a very important feature comes into play to cement the entire structure, that is *agreement.*

The Role of Agreement in MT

Agreement features are very important and should be carefully applied to ensure the generation of sound sentences in the target language. Because agreement applies to the target language, it should fulfill the specific requirements of this language. Mistakes in the MT output can be either the result of analysis problems at the source language level, or due to generation problems at the target language level.

Although word order rules are crucial for the generation of sound sentences in the target language, they are merely rules for the rearrangement of sentence constituents. These rules draw their information from the syntactic knowledge. Agreement rules, on the other hand, are more complicated. This complication becomes more obvious when making MT between languages that have great morphological variations and big differences in agreement requirements. The English verb, for example, indicates in the present tense whether the subject is singular or plural, but it does not give any information about gender. In other tenses, the English verb is completely neutral. English adjectives and determiners are also, to a great extent, neutral to number and/or gender. This poses a problem when translating English into Arabic, a language that is highly sensitive to agreement features. Arabic verbs, adjectives, and most determiners are highly reactive to the noun they modify, whether singular, dual, plural, feminine, masculine, human or non-human.

To make accurate agreement in the output, MT systems can draw the information they need in some cases from the source language. In the source language, lexical items, especially proper nouns and titles, should be defined whether they are feminine or masculine, singular or plural. We need the source language to tell us that *Jack* is masculine while *Jill* is feminine. In other cases the information needed for agreement is derived from the target language. In the target language, lexical items, especially common nouns, should be defined, whether they are feminine or masculine, singular, dual or plural. The target language will tell us that *kitāb*, which is the equivalent for *book*, is masculine while *kurrāsah*, the equivalent for *notebook*, is feminine. A more complicated case is when the information needed for agreement cannot be drawn either from the source language or the target language. In this case the context is the determining factor. The MT system makes a sort of backtracking (going

forth and backward) in search of the information it requires. For example, in the sentence '*The student likes her teacher'*, we cannot determine whether *the student* is feminine or masculine until we go forward and see that it is referred to as *her*. The gender of *teacher*, however, cannot be determined from the sentence. We have to look for clues in previous or subsequent sentences; otherwise it will take the default value of masculine.

Research Aim

The aim of this research is to explore the implications and effects of the agreement features in the MT process. The research target is to determine to what extent agreement, as a set of features and as a set of rules, is responsible for generating coherent Arabic structures in the MT output. The research will also address the issue of how the information needed for agreement can be reached when translating from a language with little morphological variations like English into a morphologically rich language like Arabic. The research will also contribute to the issue of computer knowledge acquisition, as the MT systems will be required to acquire information from the texts they are dealing with.

The Scope of the Study

The scope of this study is English into Arabic translation. I will discuss the analysis of English as a source language, problems related to the transfer of English into Arabic, and the generation of Arabic as a target language. This means that I will not discuss aspects related to the analysis of Arabic as a source language.

In MT research, there are three basic operational strategies, direct, transfer, and interlingua (not to mention knowledge-based and example-based as they are still, to a large extent, experimental). The scope of this study is the analysis of MT based on the transfer strategy. Other basic strategies will be briefly explained and referred to occasionally in the study; yet the main focus will be on transfer. My reason for this choice is that the direct strategy, as will be explained later, in one extreme, is both theoretically and practically incapable of meeting the needs of MT development, as it has no plausible theoretical background. In the other extreme, the interlingua approach is highly theoretical and is used in a limited number of systems that are not available for widespread commercial application. I believe that, in contrast to the other two, the transfer approach has proved to be both theoretically powerful and

practically available for MT application. Most MT systems available in the market today are designed according to the transfer strategy. “Transfer systems are generally regarded as a practical compromise between the efficient use of resources of interlingua systems, and the ease of implementation of direct systems.”[1]

This study will also be limited to electronic texts, i.e. texts written in machine-readable format. I will not discuss how the paper documents can be inputted to a computer. I will not discuss how the computer can analyze spoken languages either.

The Methodology of the Study

Throughout this research I will be working on MT from English into Arabic. To make the research stand on a practical ground, I will refer to the English into Arabic MT system Al-Mutarjim Al-Arabey by ATA Software. It should be noted, however, that the research does not analyze or evaluate any MT system. The MT system is used only as a testing ground for the points introduced in this paper. The research will be supported with examples from the system to back the issues discussed and show how they were actually manipulated. In some cases the MT system will be successful in making the correct agreement; in other cases it will not. In the successful instances I will show how the agreement was tackled, and in the failing instances I will show what was lacking.

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 121.

Chapter One
Theoretical Background

In this chapter I will explain the meaning of machine translation (MT), what it tries to achieve, what problems it faces, and what prospects it has. I will also give a brief history of MT as a field of study, how it began, how it developed, and what obstacles it met. Then I will explain some of the main strategies which researcher followed in the design and implementation of MT applications.

1.1. What Is Machine Translation?

Machine translation (MT) means using a computer to translate a human language into another human language without (or with minimal) human intervention. It is the attempt to make the computer acquire the kind of knowledge that translators need to perform their work. The computer needs to be provided with the appropriate procedures and routines to complete the translation process.

To successfully undertake a translation task, human translators needs to have four types of knowledge:

1) Knowledge of the source language (lexicon, morphology, syntax, and semantics) in order to understand the meaning of the source text.
2) Knowledge of the target language (lexicon, morphology, syntax, and semantics) in order to produce a comprehensible, acceptable, and well-formed text.
3) Knowledge of the relation between source and target language in order to be able to transfer lexical items and syntactic structures of the source language to the nearest matches in the target language.
4) Knowledge of "the subject matter"[1]. This enables the translator to understand the specific and contextual usage of terminology.

Ultimately, the translation process is not considered successful unless the output text has "the same meaning"[2] as the input text. Therefore, the

[1] Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 1.
[2] J. C. Catford, *A Linguistic Theory of Translation*, Oxford: Oxford University Press, 1965, p. 35.

transfer of lexical items and syntactic structures is not considered successful translation if the overall meaning is not conveyed.

In addition to the types of knowledge mentioned above, translators must have a special skill in their craft. To a great extent, translation "is an intelligent activity, requiring creative problem-solving in novel textual, social and cultural conditions."[1] Not only does the translation depend on linguistics, but it also "draws on anthropology, psychology, literary theory, philosophy, cultural studies and various bodies of knowledge, as well as on its own techniques and methodologies."[2]

It is not so easy for the computer to translate as to conduct a mathematical operation. In order for the computer to translate, it must go through three complicated barriers: the language barrier, the cross-linguistic barrier and the translation barrier. These barriers have been perplexing philosophers and linguists for ages. "In order to act upon human input the computer must be able to take it apart and form a logical representation of what it is fundamentally saying – it must to some degree 'understand' the input."[3] However, this understanding is not easily available because "human language is full of ambiguities, words and phrases that can mean several different things, shortened forms of words and sentences, and other factors that can serve to cloud meaning."[4] The meaning of a human utterance is "open to doubt, depending on such things as knowledge, context, association and background."[5] If sometimes we need our addressor to explain or paraphrase what he means, we cannot expect the computer to outsmart us in our own media of communication.

MT can never be achieved by feeding the computer with a dictionary of the source language words and their equivalents in the target language. Lexical equivalence is only a component among several components involved in MT today. After computer engineers and linguists were met with many failures in the beginning of MT application, they now understand the intricacy of the task. Many linguists and computer

[1]Douglas Robinson, *Becoming a Translator: An Accelerated Course*, London: Routledge, 1997, p. 51.
[2] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 3.
[3]Tim Willis, "Processing Natural Language", in Peter Roach, ed., *Computing in Linguistics and Phonetics: Introductory Readings*, San Diego: Academic Press, 1992, p. 51.
[4] Ibid.
[5] Marjorie Boulton, *The Anatomy of Language: Saying what we Mean*, London: Routledge & Kegan Paul Limited, 1960, p. 47.

engineers today are directing their efforts towards MT research. MT has become a "testing ground for many ideas in Computer Science, Artificial Intelligence and linguistics".[1]

Once a far-away dream, MT today has become a reality. Many advances have been made, many successes have been achieved and many translation applications are now available in the market. However, this reality is not as big as people hope. Commenting about the capacity and prospect of MT, Hutchins said, "There are no 'translating machines' which, at the touch of a few buttons, can take any text in any language and produce a perfect translation in any other language without human intervention or assistance. That is an ideal for the distant future, if it is even achievable in principle".[2] Though these words are said a decade ago, they are still expressive of the state of the art of MT today. The translation process is so complicated for the machine to handle. The machine cannot deal with all types of texts in all fields. No MT manufacturer claims that his application can produce a hundred per cent accurate and comprehensible output.

Some people argue that studies in MT are useless because the machine can never translate great literary works like those of Shakespeare or Dickens. However, translating literary works is not within the scope of MT, because "translating literature requires special literary skill"[3] and creativity from the translator. It is usually a poet or a man of letters (not a customary translator) who attempts to translate such a kind of material.

The machine cannot and will not replace translators, but it helps them in a variety of ways. MT can handle the huge routine tasks. Technical manuals and periodicals, for example, are a perfect material for MT. They use no figurative or flowery language. They have specific subject fields and restricted styles, terminology, structures, and vocabularies. MT can also provide raw translation which can be revised or 'post-edited' to give a high quality translation in a shorter time.

---

[1] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 5.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 1.
[3] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 6.

1.2. History of Machine Translation

Historically, MT passed through different stages, attitudes, and perceptions. At first there was over-exaggeration in hopes and expectations. This was in fact due to underestimation of the immensity of the task and the complexity of the problem, computational-wise to some extent, and linguistic-wise to a great extent. Then came a period of criticism and disappointment at the failure of researchers to arrive at a workable system that can deliver acceptable automatic translation. MT was distrusted as a legal field of scientific study and was considered as unachievable both in practice and in principle. The failure to arrive at a workable system at this period was due in part to hardware limitations, software limitations, and linguistic research limitations. Major developments in these areas, which were effected later, contributed to the revival of MT study until it became an attractive field of study with widespread applications. Here I will give a brief historical account of MT research.

1.2.1. First Idea

The first idea of using the computer to translate among human languages automatically, without reliance on human translators, emerged during World War II when computers were used to decipher encryptions. It is traced back to a memo by Warren Weaver in 1949 which included the following sentences:

> I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text.[1]

The Weaver's analogy of cryptography and translation is not acceptable to many linguists since coding is a "one-for-one substitution process" using different symbols for words of the same language, whereas translation is "a far more complex and subtle business"[2] involving two different structures, conceptions, and cultures. However simplistic this memo might have been, it triggered interest in MT and initiated MT research projects in the United States, Russia, and other countries.

[1] Ibid., p. 13.
[2] Ibid.

1.2.2. Widespread Optimism

By the 1950s interest in MT research has grown considerably and many MT groups were formulated in many centers in the USA and Russia. In 1952 the first MT conference was held in Massachusetts Institute of Technology (MIT) and organized by Yehoshua Bar-Hillel. In the opening session Bar-Hillel emphasized "the need and possibilities for MT, particularly to cover the immense and growing volume of scientific research and popular periodical literature."[1] Leon Dostert, from Georgetown University, suggested the creation of a pilot project to prove to the world the possibility and practicality of MT.

The conference was successful and led to high expectations. Dostert set up an MT research team in Georgetown University and cooperated with IBM to work towards the pilot project he proposed to show to the world the practical feasibility of MT. In 1954 the Georgetown-IBM experiments resulted in the first public demonstration of MT applied to Russian and English. The experiment went beyond word-for-word replacement but was limited in vocabulary (250 Russian words) and rules (6 grammar rules). However, the experiment was considered successful and "showed that MT was a feasible objective, and it undoubtedly helped to stimulate the funding of MT research by US governmental agencies in the following decade."[2] According to Hutchins[3], MT research received a massive funding in the United States: $6,585,227 granted by the National Science Foundation, $1,314,869 by the Central Intelligence Agency, and $11,906,600 by the Department of Defense. It is clear that most finance came from military and intelligence sources, the matter that indicates political motives at the time, when rivalry between the USA and USSR was getting high intensity.

Hutchins[4] states that in 1963 there were ten research groups in the United States: Georgetown, MIT, Harvard, NBS, Berkeley, Ohio State, Wayne State, Texas, Bunker-Ramo, and IBM. There were also three British groups: Cambridge, Birkbeck, and the National Physical Laboratory. There were also other strong research centers in the then USSR and Eastern Europe. All this indicated the momentum gathered by the new research area and the great expectations hanged over it. MT now

[1] W. J. Hutchins, *Machine Translation: Past, Present, Future*, West Sussex, England: Ellis Horwood Limited, 1986, p. 35.
[2] Ibid., p. 37.
[3] Ibid., pp. 167-168.
[4] Ibid., p. 167.

became a legal field of study with doctoral dissertations, scholarly journals, and books fully dedicated to it.

1.2.3. Skepticism and Setback

In the early 1960s some scientists and funding agencies were gradually starting to lose hope in MT due to the slow progress and the failure to implement an operational system. "There were not only problems of technical facilities and hardware, but also the complexities of the linguistic problem. These were becoming more and more apparent."[1]

In 1959, Bar-Hillel, one of the scientists involved in MT research since its early beginnings, published his *Report on the state of machine translation in the United States and Great Britain*. In this report Bar-Hillel agued strongly against MT and criticized the methodology and goals pursued by MT research groups at the time. He agued that "fully automatic, high quality, MT (FAHQMT) was impossible, not just at present, but in *principle*."[2] He suggested that MT should focus on moderate translation that involved human interaction.

Bar-Hillel's report cast a lot of uncertainties around MT not only in the public perception but also within MT researchers. Many ordinary people as well as scientists began to view MT as unattainable, impossible, and unachievable. The report "was held as 'proof' of the impossibility of MT. To this day, Bar-Hillel's article is still cited as an indictment of MT research… There can be few other areas of research activity in which one publication has had such an impact."[3]

Two years later, and in 1961, Bar-Hillel's report was supported by another publication by Mortimer Taube entitled *Computers and Common Sense*. In this book Taube argued that any attempt to mechanize human thinking processes, like translation, is doomed to failure. He "gave expression to a prevalent anti-computer view of the time."[4]

Upon request from the funding agencies, the National Academy of Sciences formed the Automatic Language Processing Advisory

[1] Ibid., p. 153.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 13.
[3] W. J. Hutchins, *Machine Translation: Past, Present, Future*, West Sussex, England: Ellis Horwood Limited, 1986, p. 157.
[4] Ibid., p. 162.

Committee (ALPAC) in 1964 to report on the current status of MT research and advise on its feasibility.

The committee studied the existing demand, supply and costs of translation, the availability of translators, and the cost and output quality of MT. In 1966 the committee issued its report, known as the ALPAC report, which concluded that "MT was slow, less accurate and twice as expensive as human translation."[1] The committee saw no immediate prospect for MT and so there was no need for further research and investment in this field. Instead, researchers should concentrate on "the development of machine aids for translators, such as automatic dictionaries, and continued support of basic research in computational linguistics."[2]

The APLAC report was viewed by many scholars as "narrow, biased and shortsighted."[3] Researchers protested that improvements to MT systems were possible and that the ALPAC conclusion about MT failure was premature. No matter what arguments MT defenders proposed, the ALPAC report caused a sever damage to MT study. Many researchers lost morale and the financial support of MT projects was discontinued in the United States and elsewhere for the following decade. The ALPAC report destroyed "the credibility of MT research. After ALPAC few American researchers were willing to be associated with MT."[4]

1.2.4. Revival

The revival of interest in MT started outside the United State -- in Canada and Western Europe. This is understandable in light of these countries' basic needs for translation. The "Canadian bicultural policy created a demand for English-French [translation] … and the European Economic Community (as it was then known) was demanding translations … from and into all the community languages."[5] A research group in Montreal, Canada, succeeded in 1976 in creating an English-French system (Meteo) for translating weather reports. The system,

---

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 7.
[2] Ibid.
[3] W. J. Hutchins, *Machine Translation: Past, Present, Future*, West Sussex, England: Ellis Horwood Limited, 1986, p. 167.
[4] Ibid., p. 169.
[5] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 7.

though limited in its scope, was successful and practically useful. In the same year the Commission of the European Communities (CEC) purchased the English-French version of the Systran system, developed by the Georgetown team. Systran was originally a Russian-English system used by both the United State Air Force and NASA. The CEC also requested the development of English-Italian, English-German and other language pairs in the system. In the late 1970s the CEC started to fund work on the Eurotra system. This was an ambitious project which aimed at the development of a multilingual interlingua[1] system for all the community languages.

The projects which started in the 1970s and 1980s proved to be practically and commercially successful. This indicated that MT was "firmly established, both as an area of legitimate research, and a useful application of technology."[2]

1.2.5. Large-Scale Application

In the late 1990s and the start of the new millennium we saw useful and powerful MT systems on personal computers and on the Internet. Major enhancements to the MT systems have been implemented both in speed and performance. Up to this day the fully automatic high quality machine translation (FAHQMT) dream has not been realized, but MT output is useful in giving the reader a gist of the article he is interested in, or as a translation draft that requires post-editing to get a reliable translation.

[1] An approach to MT, explained in the next chapter.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 16.

1.3. Basic Machine Translation Strategies

Different strategies were adopted by different research groups at different times. Strategy choice reflects both the depth of linguistic manipulation and the breadth of ambition. At the early stages of MT research and development, little was understood about linguistic complexities. A simple methodology was followed by replacing source language (SL) words with their equivalents in target language (TL) with few rules for local reordering. As MT research grew, scientists concentrated more on the analysis of SL with higher levels of abstractness. When the aim was to translate among a large combination of language pairs, the level of abstractness of language analysis went even higher. In this section I will explain these strategies in detail.

1.3.1. Direct

The direct strategy stands for the approach where "the processing of source language input text leads 'directly' to the desired target language output text"[1] without any intermediate analysis. This strategy was characteristic of almost all systems developed in the 1950s and 1960s, which were known as the 'first generation' of MT systems. "A direct translation system is designed, from its outset, for a specific source and target language pair. No general linguistic theory or parsing principles are necessarily present."[2]

Processing in MT systems that follow the direct strategy consists of three stages:

1. Morphological analysis of the source language input text. In this stage the system identifies word ending and reduces inflected forms to their uninflected base forms.
2. Bilingual dictionary lookup. Depending on a huge bilingual dictionary the system decides the correct replacement for source words with equivalent words in the target language.
3. Local reordering of the target language. After the replacement is done the system makes adjustment to the output text by applying rules for

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 72.
[2] Allen B. Tucher, "Current strategies in machine translation research and development", in Sergei Nirenburg, ed., *Machine Translation: Theoretical and methodological issues*, Cambridge: Cambridge University Press, 1987, pp. 22-23.

putting words in their right order. The following figure summarizes this process.

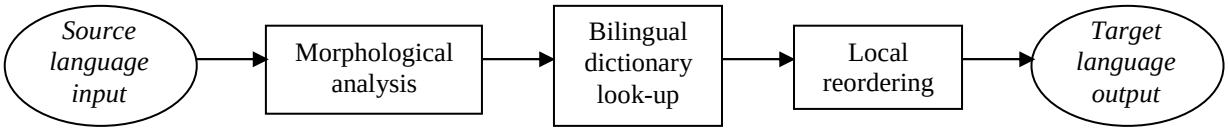


Fig. 1. Direct MT strategy[1]

As can be seen from fig. 1 above, no analysis of syntactic structures or semantic relationships is applied.

Computational and linguistic simplism of this strategy is obvious. MT systems were developed by primitive computers with less speed, performance, and processing capability than the cheapest PCs of today. Programming was done in assembly code, not in the high-level programming languages used today. The output was only "word-for-word" translation; the kind of translation that can be expected from a person with simple knowledge of the target language attempting to translate a text using only a bilingual dictionary.

Despite the clear disadvantages of the direct method, it is still applied to some extent in unidirectional bilingual systems. These systems "take advantage of similarities of structure and vocabulary between source and target languages in order to translate as much as possible according to the direct approach"[2].

1.3.2. Interlingua

The failure of the direct MT systems, or what is called 'the first generation', led the scientific research into the development of indirect MT systems, or what is called 'the second generation'. However, there are two variants of the indirect method depending on the degree of analysis of the source and target languages and the need for comparative grammar and transfer rules.

Historically speaking, the first indirect method is the interlingua where the source language is analyzed in an intermediate "semantico-syntactic"[3]

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 72.
[2] Ibid., p. 73.
[3] W. J. Hutchins, *Machine Translation: Past, Present, Future*, West Sussex, England: Ellis Horwood Limited, 1986, p. 54.

representation from which the target language is generated. This intermediate representation is neutral in the sense that it does not carry features either from the source or target languages. “In the past, the intention or hope was to develop an interlingual representation which was truly ‘universal’ and could thus be intermediary between any natural languages. At present, interlingual systems are less ambitious.”[1]

The basic idea behind the interlingua method is that the deeper the linguistic analysis goes the less becomes the need for comparative grammar and transfer rules. In interlingual systems translation is carried out in two stages. In the first stage the source language is analyzed into interlingual representation and in the second stage the target language is generated from this representation. The interlingual representation “would have to be entirely language independent”[2] in the sense that it does not carry any specific features of the source language and is not designed with any specific target language in mind. However, researchers do not agree on the nature of this intermediate interlingual representation: “a ‘logical’ artificial language, or a ‘natural’ auxiliary language such as Espranto; a set of semantic primitives common to all languages, or a ‘universal’ vocabulary, etc.”[3]

There is a clear advantage of the interlingual method. It is easier to add new language pairs to the system than in the transfer method, as “the addition of a new language to the system entails the creation of just two new modules: an analysis grammar and a generation grammar.” For example, in a system that has four languages (English, French, German, and Spanish) there are 12 language pairs:

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 73.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 80.
[3] W. J. Hutchins, *Machine Translation: Past, Present, Future*, West Sussex, England: Ellis Horwood Limited, 1986, p. 55.

1)	English	=>	French
2)	English	=>	German
3)	English	=>	Spanish
4)	French	=>	English
5)	French	=>	German
6)	French	=>	Spanish
7)	German	=>	English
8)	German	=>	French
9)	German	=>	Spanish
10)	Spanish	=>	English
11)	Spanish	=>	French
12)	Spanish	=>	German

This can be achieved only by the creation of eight modules: four for each language analysis and four for each language generation, as can be seen in the following figure:

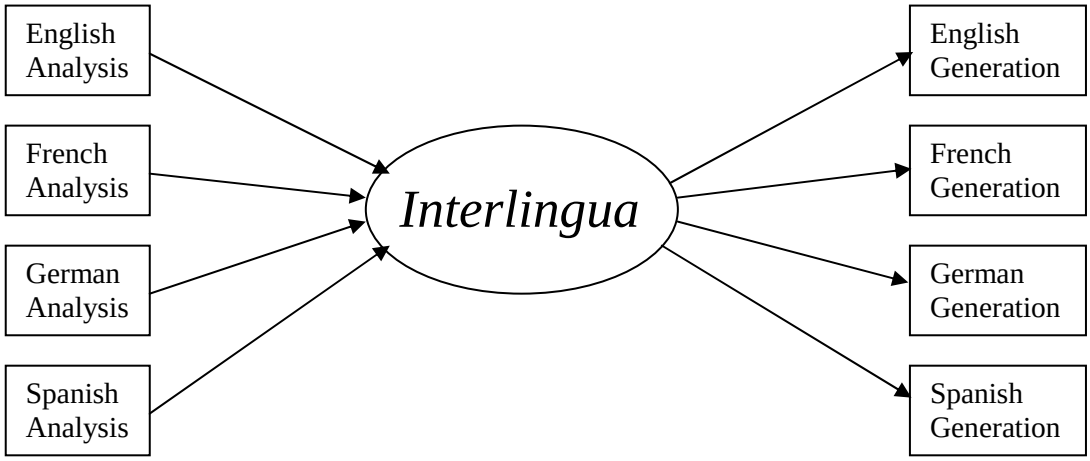


Fig. 2. Interlingua MT with 12 language pairs[1]

It can be noted that in the interlingua model it is possible to translate from and into the same language. This can be achieved by translating from English, for example, into the interlingua and then back into English again. The output will not be the same as the input but a paraphrase or summary of it. "This seemingly 'back translation' capability could in fact be extremely valuable during system development in order to test analysis and generation modules."[2]

A major disadvantage of the interlingua method is the difficulty to design an interlingua even for closely related language. "A truly

[1] Adapted from W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 74.
[2] Ibid.

'universal' and language-independent interlingua has defied the best efforts of linguists and philosophers from the seventeenth century onwards."[1] This makes it extremely difficult to produce an operational system based on a genuine interlingua.

1.3.3. Transfer

The transfer method is a middle course between direct and interlingua MT strategies. The difference between the three strategies can be captured in the following figure:

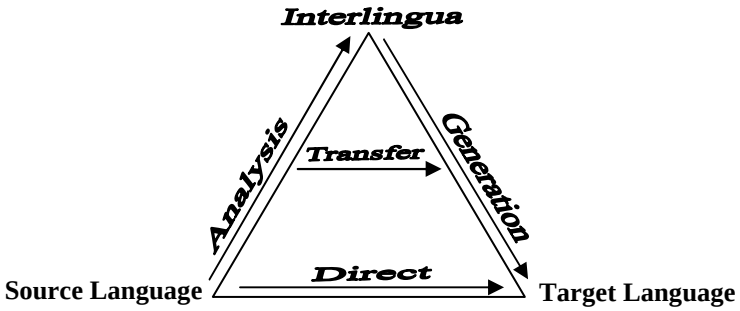


Fig. 3. Difference between direct, transfer, and interlingua MT methods[2]

As can be seen from the above figure, the direct method has no modules for source language analysis or target language generation but applies a set of rules for direct translation. In interlingua method the source language is fully analyzed into a language-independent representation from which the target language is generated. The transfer method cuts the road in the middle. The transfer strategy can be viewed as "a practical compromise between the efficient use of resources of interlingua systems, and the ease of implementation of direct systems."[3] The source language is analyzed into a language-dependent representation which carries features of the source language. Then a set of transfer rules are applied to transform this representation into a representation that carries features of the target language. At the end the generation module is used to produce the target output.

There are two advantages of the transfer method that make it appealing for many researchers:

---

[1] Ibid., p. 75.
[2] Adapted from Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 6.
[3] Ibid, p. 121.

1. Applicability. While it is difficult to reach the level of abstractness required in interlingua systems, the level of analysis in transfer models is attainable.
2. Ease of implementation. Developing a transfer MT system requires less time and effort than interlingua. This is why many operational transfer systems have appeared in the market.

One clear disadvantage of the transfer method is that it is costly when translation between many languages is required. The transfer method "involves a (usually substantial) bilingual component, i.e., a component tailored for a specific SL-TL pair."[1] This entails significant effort and time for each new language added to the system. If you develop a system that translates among four languages, you will need 12 transfer modules in addition to four analysis and four generation modules. This makes a total of 20 modules. Mathematically speaking, the number of transfer modules for *n* languages is "$n \times (n - 1)$"[2] in addition to *n* analysis and *n* generation modules. The following figure shows the complexity of the task when designing a multilingual transfer system.

[1] Allen B. Tucher, "Current strategies in machine translation research and development", in Sergei Nirenburg, ed., *Machine Translation: Theoretical and methodological issues*, Cambridge: Cambridge University Press, 1987, pp. 23-24.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 76.

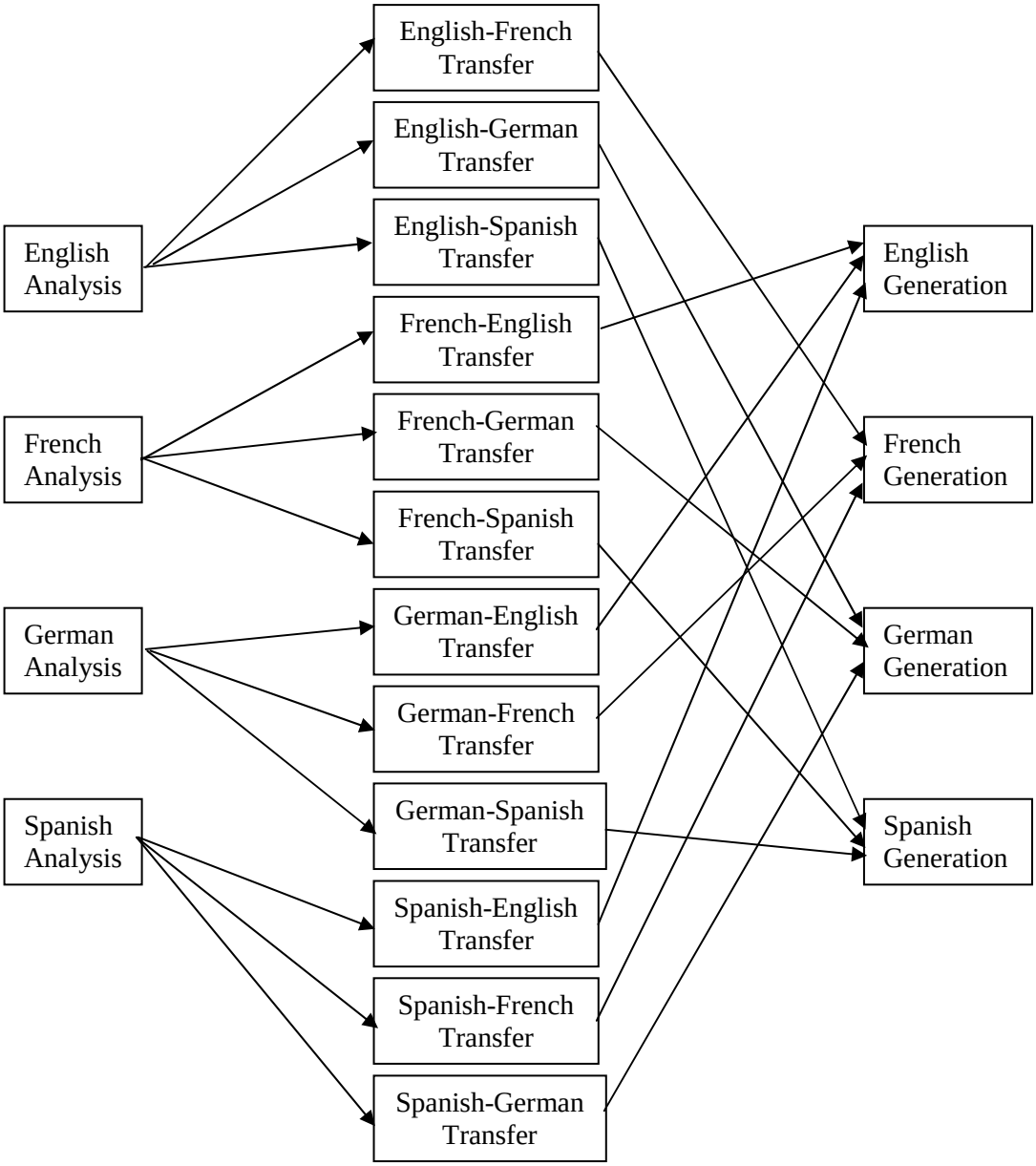


Fig. 4. Modules in a transfer MT system with 12 language pairs[1]

By comparing the above figure with figure 2 we can clearly see the advantage of interlingua design over transfer when a number of languages are included in the system. The interlingua method is more economical because it dispenses with transfer modules by analyzing source languages into language-independent interlingual representation from which target texts are directly generated.

However, the picture is not as gloomy as it appears. It is not impossible to develop a multilingual system using the transfer method. "If the design is optimal, the work of transfer modules can be greatly simplified and the creation of new ones can be less onerous than might be

[1] Ibid. (adapted).

imagined."[1] There are a number of techniques that can be applied to improve the cost-effectiveness of the transfer method in a multilingual environment:

1. Making deeper analysis of the source language. As it has been noted, the deeper the analysis goes the less work is required in the transfer component.
2. Making use of reversibility. If a transfer module from English into French for example can be reversed this will reduce the work required by half. However, it should be noted that not all transfer rules in a component could be reversed. "The transfer rules could be reversible *in principle*, and though this is natural, and attractive …, it is not obvious that reversible transfer rules are always possible or desirable."[2]
3. Sharing some transfer rules with closely related languages. "Portions of transfer modules can be shared when closely related languages are involved. For example, an English-Portuguese module may share several transformations with an English-Spanish module."[3] In this way some transfer rules are reusable and can be shared by different transfer components.

1.3.4. Knowledge-Based MT

Knowledge-Based MT (KBMT) systems are based on the fact that "high quality translation requires in-depth understanding of the text"[4]. Translation requires reference to the real world knowledge in addition to knowledge of the "differences in cultural backgrounds and differences in conceptual divisions"[5] among different languages. The word "rice" in English, for example, has six different translations in Malay depending on whether it is harvested or not, cooked or not. This is an instance of the difference in cultural backgrounds. The word "wear" in English has eight different meanings in Japanese depending on the object to be worn. This is an instance of conceptual divisions.

---

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 76.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, pp. 78-79.
[3] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 122.
[4] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 190.
[5] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 124.

In KBMT systems, translation is based on non-linguistic, interlingual conceptual representation of "meanings derived from the processes of **understanding** of texts."[1] These interlingual representations can serve as intermediate representations, and with appropriate knowledge-bases, texts can be analyzed and generated from these representations. KBMT systems rely on information that cannot be derived from linguistic inputs alone, but include real world knowledge. As a result, "KBMT systems rely on an augmentor"[2]. The augmentor helps the knowledge bases by adding more information to them. Augmentation can be done by the machine through inferences about the input text or by humans interacting with the machine and providing needed information. A model of KBMT system can be pictured in the following figure.

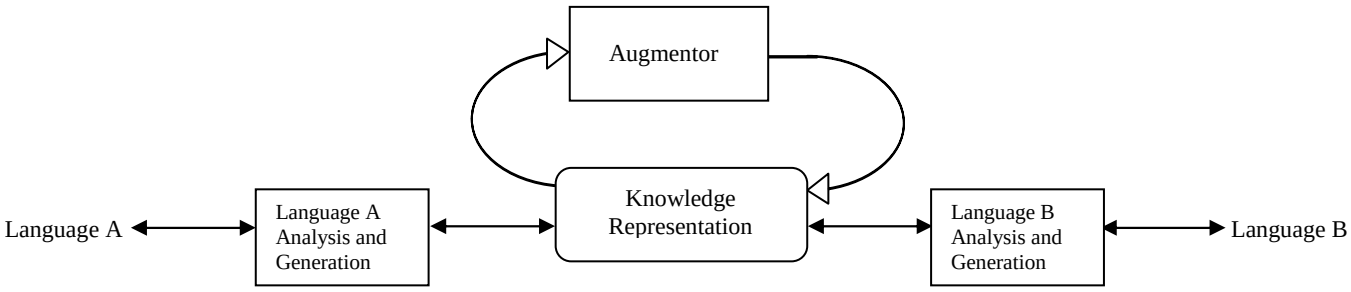


Fig. 5. KBMT architecture[3]

Knowledge bases try to simulate the knowledge a translator would have. They are difficult to build and maintain, and this is why KBMT systems are highly restricted to specific domains and sublanguages "with relatively narrow contexts and applications"[4] like technical reports and manuals.

1.3.5. Example-Based MT

All of the previous methods (direct, interlingua, transfer, and knowledge-based) are classified under rule-based approach. They state explicit rules for manipulating the translation process. Another approach is corpus-based which tries to manipulate the translation process by

[1] Ibid., p. 125.
[2] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 182.
[3] Ibid., p. 183.
[4] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 125.

referring to a corpus of previous translations. A clear instance of this approach is the Example-Based MT (EBMT) systems. The idea behind this system "is to collect a bilingual corpus of translation pairs and then use a best match algorithm to find the closest example to the source phrase in question. This gives a translation template, which can then be filled in by word-for-word translation."[1] It is based on the idea that new translations are often modifications of previous ones and that a good translator always refers to previous translations to save time and ensure consistency in terminology and style.

If the input sentence has an identical match in the corpus, the system retrieves the translation with no further manipulations. But this is not always the case as identical matches are very rare (due to the nature of human languages). If the sentence has no identical match the system analyzes it into parts and match these parts against parts in corpus sentences. The matched parts are then combined to form translation for the new sentence. However this process is not as easy and straightforward as it appears. It is difficult for the system to identify the translation for a specific part within a sentence. "Algorithm for word alignment may be used to address this problem, but error rates are still significant."[2] Furthermore, translation is not one-for-one match; it may involve additions, deletions, or inversion of structures. This is due to the divergences and mismatches among languages.

Due to the complexities involved in EBMT, it is not recommended to be used as a stand-alone translation solution, but as a supplementary aid for rule-based systems. It is "evident that example-based approach can be integrated in any of the basic models: direct, transfer, and interlingua."[3]

---

[1] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 198.
[2] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 204.
[3] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 127.

Chapter Two
Machine Translation Processes

This chapter focuses on the procedures involved and requirements needed in the MT process. It will explain how the input source text is analyzed and processed until at the end the output source text is generated. The process of source text analysis goes through many steps. It starts with the minimal level of representation, a single alphabet letter or character. The second level is the word at the morphological level. Then comes the dictionary, or bilingual lexicon. Next to this is the analysis at the sentential level, which is dealt with in terms of syntax and phrase-structure trees. Each of these levels has its problems and ambiguities on which linguists and MT specialists have concentrated their attention in order to provide appropriate solutions.

2.1. Analysis of the Source Language

The first problem faced by MT developers is the analysis of the source language text. Analysis "concerns the application of monolingual rules to the source language input"[1] in order to recognize the structure of the sentence. Analysis does not occur at a single level, but at different hierarchical levels. The MT system has first to recognize the electronic format or characters of the input text, then it goes up to the word level, or morphology; the meaning level, or semantics; then the sentence level, or syntax. Breaking down the analysis into several levels has a great advantage: "Not only does this break down the problem into more manageable components, but it allows each level to develop its own methodology."[2] Linguists and programmers do not have to deal with the analysis as one big problem, but they deal with it through small manageable modules. Any problem that later emerges can be solved only within the related component. Each level can also develop its own methodology independently.

It must be noted, however, that the levels of analysis are not agreed upon by all researchers, as there is no "standard set of levels used by all

[1] Frank Van Eynde, "Machine Translation and Linguistic Motivation", in Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 73.
[2] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 10.

linguists."[1] What I will present here is a global framework of the analysis problem that must be tackled by any MT system.

Another issue of major importance here is the relation between computational linguistic methodologies and theoretical linguistics in general. Although computational linguists take their basic ideas from theoretical linguists, there are some points that must be made clear:

1. MT does not follow any theory strictly. No theory in theoretical linguistics reached the level of descriptive adequacy needed in MT. So, even if some MT developers claim to follow a certain theory, they later find themselves obliged to break with the theory somewhere.
2. MT does not depend on a single linguistic theory. Although some theories (like Transformational Grammar, Generalized Phrase-Structure Grammar, and Dependency Theory) have greater influence in MT, "the great majority of MT systems are amalgams of different approaches and models, or even occasionally … with no discernible theoretical basis at all."[2] MT researchers borrow from different theories what they find adequate for meeting their immediate needs.
3. MT takes a more pragmatic approach to linguistics. It is a kind of "engineering"[3] which looks more at problem-solving methods. The primary interest in theoretical linguistics is still to answer the abstract questions of language faculty, use, and acquisition. This can provide little help to MT whose primary interest is to give real answers and practical solutions.
4. Most researchers in theoretical linguistics concentrate on English. They do not try to capture differences between languages. They rarely address languages contrastively or describe how languages use different means to express the same concepts. "Such questions are of course at the heart of MT."[4] In order for MT systems to work, two languages at least (source and target languages) must be fully analyzed. Moreover, similarities and dissimilarities between these two languages must be explored in detail. This kind of study, however, is not adequately covered by theoretical linguistics.

---

[1] Ibid.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 81.
[3] Ibid., p. 82.
[4] Ibid.

2.1.1. Encoding and Character Sets

The first problem an MT system faces is how to read the characters representing the input language. English characters are different from Arabic, German, Chinese and even French characters. The system must be able to recognize the input language and process the characters correctly.

Due to the fact that "computer technology has been mostly developed in the USA, one of the most thoroughly monolingual societies in the world,"[1] the introduction and processing of writing systems other than English posed a considerable problem. This problem was especially felt in the field of translation where more than one language was involved. In 1997 when I was working in the Translation Department in Harf Information Technology, we faced a big problem regarding Arabic character sets. We were working with applications from three different companies: Sakhr's Arabic Windows, IBM's Translation Manager 2, and Microsoft Windows 3.x with Arabic Support. Each company was using a different encoding system for Arabic. This caused considerable confusion and we had to devise a tool for conversion among the different systems.

Electronic character sets are groups of characters represented in ASCII codes. It was used for English in 1963. "This standard defines 127 codes (a 7-bit standard) which included the Latin characters used in English, together with a number of control, punctuation, and symbol characters."[2] The English 26 alphabet letters are represented by 52 ASCII codes, with special encoding for upper and lower case letters. The ASCII code for the capital letter *A*, for example, is 65. However, this ASCII standard could not be used for displaying several European languages as well as Arabic. "This led to the introduction of the ISO-8859 series of 8-bit standards."[3] The ISO standard 8859-6 was allocated for Latin/Arabic character sets. Arabic requires the representation of 44 different characters. These include the 28 alphabet letters as well as 8 diacritic marks and varieties of *hamza*'s and *yaa*'s. Arabic letters and their ASCII codes are listed in the following table:

[1] Ibid., p. 52.
[2] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 47.
[3] Ibid., p. 48.

193 : ء	194 : آ
195 : أ	196 : ؤ
197 : إ	198 : ئ
199 : ا	200 : ب
201 : ة	202 : ت
203 : ث	204 : ج
205 : ح	206 : خ
207 : د	208 : ذ
209 : ر	210 : ز
211 : س	212 : ش
213 : ص	214 : ض
216 : ط	217 : ظ
218 : ع	219 : غ
221 : ف	222 : ق
223 : ك	225 : ل
227 : م	228 : ن
229 : ه	230 : و
236 : ى	237 : ي
240 : ً	241 : ٌ
242 : ٍ	243 : َ
245 : ُ	246 : ِ
248 : ّ	250 : ْ

Table 1. Arabic letters and their ASCII codes

Although this code convention accommodates most languages, there are still some disadvantages. It is extremely difficult to mix texts from different standards; that is, you cannot mix Arabic and German, for example. Furthermore, languages like Japanese and Chinese do not use alphabetical writing systems, but thousands of different characters. Japanese has 3,000 characters and Chinese 6,000. No expansion to ASCII can ever solve this problem. So the computer industry developed what has come to be known as the Unicode standard.

The Unicode standard is “a 16-bit (2-byte) code developed by the Unicode Consortium of major software developers and other interested parties.”[1] This new standard accommodates 65,536 different characters, which are enough to represent all the languages of the world. Each

[1] Ibid., p. 49.

character is represented by a unique hexadecimal number. This allows computers to deal with any mix of languages without any confusion.

Another peculiarity to the Arabic writing system (besides characters) is direction. Whereas most world languages are written and read from left to right, Arabic and Hebrew are the only known languages to have right-to-left direction. This issue cannot be ignored in any MT system dealing with Arabic either as input or output.

2.1.2. Orthography

Orthography covers the aspects of spelling, type style, and punctuation. Spelling is concerned with "the knowledge of possible combinations of letters"[1] in a language. This knowledge can be useful when the system encounters a misspelled word. In this case the MT system has one of two options: either to suggest a correction of the misspelled word or to leave it untranslated and deal with it as an unknown word. Most MT systems assume that the input document passed through a pre-editing phase and free of spelling errors, and so they deal with misspelled words as unknown words.

The other aspects are the problems of type style, capitalization, and punctuation. Type style means whether a word is in bold, italics, or underlined. Capitalization means whether a word is in lower case (word), upper case (WORD) or title case (Word). 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 itself. There are many orthographical differences between languages, particularly Arabic and English, which have to be taken into consideration by MT developers.

English uses italics to indicate emphasis to a word, but emphasis in Arabic is indicated by a change in word order or the introduction of an emphatic word. English proper nouns start with a capital letter, but Arabic does not give any special marking for proper nouns. Furthermore, punctuation marks in Arabic are not used so regularly as they are in English. Therefore, during the process of translation, the system must be able to identify which punctuation marks are redundant in the other language.

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 14.

The following sentence, for example, uses commas abundantly. Hardly any of these commas will be transferred in the translation into Arabic.

This year, the man, however, and his wife, too, will go on holiday.[1]

2.1.3. Morphological Analysis

Morphological analysis is one of the most basic components in any MT system. "Morphology is concerned with the way in which words are formed from basic sequences of phonemes."[2] MT systems rely on the morphological component for the analysis of words and understanding the relationship between the different forms which a single word can take. Natural language applications, such as spelling checkers, electronic dictionaries, information retrieval systems, and MT systems, need morphological analysis in order to identify newly-formed words and to ensure that words which are inflectional variants of each other are treated the same. "Analysing and generating word forms is a crucial step in the processing of natural language… NLP and MT systems need to identify words in texts in order to determine their syntactic and semantic properties."[3] In this way morphology is regarded by many MT specialists as a means of simplifying the problems of lexical analysis and a prerequisite for syntactic and semantic analysis.

2.1.3.1. Inflection, Derivation, and Compounding

Morphology is traditionally classified into three main spheres: inflection, derivation, and compounding. Inflectional morphology "deals with the formation of different forms in the paradigm of a lexeme."[4] It is concerned with the way words reflect grammatical information, and this is why some linguists tend to refer to the morphemes related to this category as "grammatical morphemes"[5]. In inflectional morphology

---

[1] Ibid.
[2] Ibid., p. 15.
[3] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, pp. 85-86.
[4] Paul Bennet "The interaction of syntax and morphology in machine translation", in Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 73.
[5] Ronald W.Langacker, *Language and its Structure: Some fundamental linguistic concepts*, 2nd ed., New York: Harcourt Brace Jovanovich, 1973, p. 76.

words undergo a change in their form to express some grammatical functions but their syntactic category remains unchanged.

Many inflectional features appear on words to express agreement purposes (agreement in person, number, and gender) as well as to express case, aspect, mood, and tense. This way, morphology is said to carry out some syntactic functions. "Frequently … morphological analysis cannot be divorced from syntactic analysis."[1] Here is a look at the morphological inflections which carry syntactic meanings (as classified by Trujillo[2]):

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

Number: It has four main contrasts: singular, dual, trial and plural. English makes 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 is given grammatical properties accordingly.

Case: It "indicates the role of a participant within a phrase."[3] Its typical contrasts are nominative, accusative, genitive, and partitive. Nominative is the case of the subject of a finite verb, e.g. *I* in *I wrote the letter.* Accusative is the case of the direct object of verb or preposition, e.g. *me* in *He talked to me.* Genitive is the case that expresses possession, e.g. *his* in *This is his car.* Partitive is the case that indicates a part as distinct from a whole, e.g. *some of the apples* in *He took some of the apples.* There are also other cases like dative, which expresses the recipient of an action; vocative, which expresses the person being called or addressed; and locative, which expresses place or time at which an action takes place.

Tense: It expresses whether an action is performed in the past, present, or future.

Aspect: It expresses whether an action is complete (*He has gone*), progressive (*He is reading*) or habitual (*He wakes up late every day*).

---

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 83.
[2] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, pp. 11-12.
[3] Ibid., p. 12.

Mood: It expresses factuality, likelihood, possibility or uncertainty. Its typical contrasts are: indicative (*He went to the school*), subjunctive (for hypothetical actions: *If he studied, he would succeed*), imperative (to express a command: *Open the door!*), interrogative (to ask a question: *How are you?*), exclamatory (to express surprise: *What a car!*), and optative (to indicate a wish: *Had I the means, I would travel abroad*).

Voice: It expresses the relation of the subject to the action. Its typical contrasts are active (*I teach*) and passive (*I am taught*).

The second sphere of morphology is derivational morphology, which is concerned with "the creation of a new lexeme via affixation."[1] In English, the process of word formation through derivation involves two types of affixation: prefixation, which means placing a morpheme before a word, e.g. *un-happy*; and suffixation, which means placing a morpheme after a word, e.g. *happi-ness*.

Derivation poses a problem to translation in that "not all derived words have straight-forward compositional translation as derived words."[2] In English, for example, the same meaning can be expressed by different affixes. Take, for instance, the nominalization process, which can employ *-ance* as in *acceptance*, *-ment* as in *development* or *-ation* as in *determination*. Moreover, the same affix can have more than one meaning. This can be exemplified by the suffix *-er*. This suffix can be used to express the agent as in *player* and *singer*. But this is not the only meaning it can convey as it can describe instruments as in *mixer* and *cooker*. In this way the affix can have a range of equivalents in the target language and the attempt to have one-to-one correspondences for affixes will be greatly misguided.

While English employs affixation in derivation, Arabic does not follow this technique. Instead it uses lexicalization, which is the formation of lexical items, to make derivatives. When no lexical item is found in Arabic to match the new meaning which the affix adds in English, a compound or even a syntactic structure is formed. This can be shown by the following examples:

---

[1] Paul Bennet "The interaction of syntax and morphology in machine translation", in Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 73.
[2] Ibid., p. 78.

1. In English the suffix *-er* is added to a verb to indicate the agent, while Arabic forms a word for this purpose:

play	player
laʿiba	lāʿib

2. In English the prefix *un-* is added to an adjective to negate its meaning, while in Arabic a compound is used for this purpose:

happy	unhappy
saʿīd	ġairu saʿīdin (or laisa saʿīdan)

3. In English the prefix *re-* is added to a verb to indicate repetition, while Arabic uses a syntactic structure:

play	replay
laʿiba	laʿiba marratan uḫrā (or ʾaʿāda-l-laʿib)

The third sphere of morphology is compounding, which is the process of forming a new word through combining two or more words. "In English, terms are often coined by the simple juxtaposition of nouns… in the course of time, some juxtaposed nouns may be fused and become a single noun."[1] The first problem which an MT system encounters with compounds is a problem of identification. The system must recognize that two words or more constitute a single compound, and must not deal with them as separate words. In English compound nouns can have one of three forms: two words fused together as in *horseback*; two words joined by a hyphen as in *horse-trade*; or two juxtaposed words separated by space as in *horse doctor*. The compound noun can pose a problem in translation regarding the compound meaning. "The meanings of compounds are sometimes obvious from their components (*blackberry*), sometimes slightly different (a *blackboard* is a special type of board, typically but not necessarily black) and sometime completely opaque (a *blackleg* is a traitor or strike breaker)."

An important notion in compounding is the notion of head. A compound noun is divided into head and modifier or modifiers. Take, for instance, the compound noun *watchtower*, which can be represented as a head and modifier:

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 84.

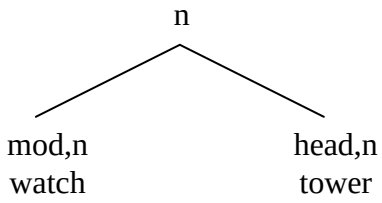


Fig. 6. A tree of English compound noun

The head does not only determine the category of the compound, but it also "determines its semantic features and (in some languages) its gender."[1] This is applicable in Arabic where the gender and number of the compound noun is determined by its head. We can look at the representation of the Arabic equivalent for the previous compound, *burǧu-l-murāqabah*:

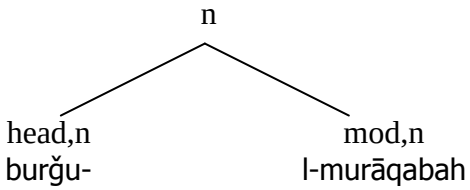


Fig. 7. A tree of Arabic compound noun

The example above sheds light on three main differences between English and Arabic compounds:

1. An Arabic compound is composed of two separate words while an English compound can be composed of two joined, hyphenated, or separated words.
2. The head, or governor, in an Arabic compound precedes the modifier while in English the head follows its modifier. The notion of head here is helpful because it "neutralizes ordering differences between languages which differ in the position of the head in compounds."[2]
3. In Arabic the head of the compound determines its gender, which in the above example is masculine, while in English this information is not relevant.

It is impossible with compounding to find a straight-forward compositional translation, as a compound structure varies from one

[1] Ibid., p. 90.
[2] Ibid.

language to another. We can now look at some of the points which show the differences:

1. Sometimes, the single noun which functions as a complement in a compound must be translated as a plural. Compare the following English-Arabic examples:

   book trade
   tiğāratu-l-kutub
   [trade of books]

2. English can use bound morphemes in compounding while other languages, including Arabic, cannot. Compare the following English-Arabic examples:

   eco-catastrophe
   kāriṯatun bīʾiyyah
   [ecological catastrophe]

3. The structure NN compound can "be translated as relational adjective + noun"[1]. Compare the following English-Arabic examples:

   atom bomb
   qunbulatun ḏarriyyah
   [atomic bomb]

4. Sometimes a compound is not translated by a compound but by a single lexical item. Compare the following English-Arabic examples:

   looking glass
   mirʾāh

   palm tree
   naḫlah

5. Sometimes an NN compound is not translated by a compound of the same structure, but by a syntactic structure of N + preposition + N sequences. Compare the following English-Arabic examples:

   team member
   ʿuḍūwun bi-l-farīq
   [member in the team]

or the opposite:

   prisoner of war
   ʾasīru ḥarb

[1] Ibid., p. 94.

Some of the key notions in dealing with morphology in MT are: productivity, generality and lexicalization. “A morphological process is said to be productive if it can be used in the formation of new words. A process is said to be generalized to the extent that it can be seen occurring in existing words.”[1] To make the distinction between generality and productivity clear, we can look at the suffix *-ment* which is used to form nouns as in *development*. This is said to be generalized as it occurs in many words, yet it is not productive as it cannot be used in new words. Yet the prefix *un-* is said to be both generalized and productive as it can be used in new words like *un-Egyptianized*.

Productivity means that new words can be formed either by adding a prefix, suffix, or using it in a compound. In the above example, *un-Egyptianized*, we can notice the productivity of derivation and inflection. The adjective denoting nationality *Egyptian* was changed into verb by adding the suffix *-ize*, and then changed into adjective by using the past participle form, and then negated by using the prefix *un-*. Compounding is also highly productive in the sense that new words can be formed freely by combining two words, which then can be extended by combining with another noun as in *information retrieval systems*. “A novel compound creates a problem for morphological analysis in an MT system: to treat it as an unknown word is unrealistic, since its meaning and the correct translation can often be derived from its component parts.”[2] So, an MT system cannot rely merely on its dictionary to identify words and reach their correct translation. It must make account for newly-formed words and ensure that it has the proper component to deal with them. “It is productivity that necessitates the treatment of morphology in MT systems. If new words can be formed, words cannot all be entered in dictionaries.”[3]

The second essential concept in morphology is generality, which means that a certain morphological process is repeated in a large amount of words, or a certain morpheme can attach itself to many words to express a specific meaning. MT systems can exploit the generality feature to reduce the size of dictionary and capture the similarity of meaning conveyed by morphological processes like derivation and compounding.

---

[1] Ibid., p. 75.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 84.
[3] Paul Bennet “The interaction of syntax and morphology in machine translation”, in Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 76.

"A generalized process will be responsible for a mass of words (e.g. noun plurals, adverbs in *-ly*, adjectives prefixed with *un-*). It may be theoretically possible to list all such forms, but such a task would be enormous,"[1] and would fail to capture the regular meanings conveyed by the morphological processes.

Lexicalization means that the new form of the word is "neither productive nor generalized."[2] This may be exemplified by the word *warmth*, which is the noun of *warm*. Here the noun was created by suffixing *-th*. However, this process is not a regular morphological process and cannot be used in new words; it is not productive or generalized, and so it is regarded as lexicalized.

2.1.3.2. Methods of Morphological Analysis

A major decision to make in the design of an MT system is to decide whether to use a morphological analysis component to analyze words and identify them as roots and affixes, or to use a database to store all full-form words. These are the two main techniques in dealing with morphology in MT systems. Each of these techniques has its advantages and disadvantages. Nonetheless, using a morphological analysis component is linguistically motivated and theoretically more attractive.

Using full-form words means that the root and all its derivatives are stored in a database. Each of the items *play*, *played*, *playing*, *plays*, *replay*, *replayed*, *replaying*, *replays*, *player*, *players*, *playful*, *un-playful*, *playfully*, *un-playfully*, *playfulness* and *un-playfulness* will all be entered explicitly into the database to be identified as relating to the same root *play*. Using this technique in dealing with morphology in an MT system has some advantages. It "makes access to words faster"[3] as the system will reach the word directly instead of going through a component for morphological analysis. It also avoids the "perceived difficulties of dealing with irregular forms and the wide range of possible inflectional paradigms."[4] When the system uses a full-form database it will not have to bother about irregular forms, as all words are treated in the same way and entered explicitly.

---

[1] Ibid.
[2] Ibid.
[3] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 86.
[4] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 82.

This technique, however, has strong disadvantages. It fails to capture regularity in the morphological process, and the task of entering all word forms becomes even more tedious with highly inflected languages like Arabic. It also fails to identify newly-formed words. As I mentioned above, morphology is productive in the sense that it can apply certain processes to create new words. If a system cannot understand these processes, it will not be able to deal with new words. Another disadvantage is that it has a greater updating cost. If a modification is required, it has to be entered in all words in the database.

Contrastively, a morphological analysis component is a rule-based module which is able to analyze a word and relate it to its root form and interpret meaning chunks conveyed by the affixes attached to the word. The component deals with a word as composed of root and affixes. This is why it is sometimes called a "two-level model of morphology"[1]. Its advantages and disadvantages are the opposite of the full-form technique. It can capture morphological generality and identify newly-formed words. The cost of updating and maintaining the system is minimal as modification is made in a single module, which is then applied to all words. Yet the disadvantages are that the cost of developing and maintaining morphological rules may be higher than that of the first technique. It may also take a longer time during processing. "In addition, irregular forms such as *be*, *am*, *are* and *is* still need to be entered separately."[2] Yet this model is the one recommended by linguists and MT specialists, as its advantages outweigh most of its disadvantages.

2.1.3.3. Lexical Databases

A lexicon provides "the specific information about each individual lexical item (word or phrase) in the vocabulary of the language concerned."[3] While grammatical rules define possible grammatical structures in a language, a lexicon states which words can appear in which constructions. It is expected that lexicons "contain all the 'idiosyncratic', 'irregular', or unpredictable information about words."[4]

---

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 86.
[2] Ibid.
[3] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 78.
[4] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 91.

MT lexicons have a great importance in MT research and design for the following reasons:

1. Lexicons are used in morphological, syntactic and semantic processing as well as in transfer. They provide information necessary for other components.
2. Lexicons are the largest and most expensive components in MT systems. They hold a large amount of data and they take a considerable time to construct. To give a hint at the number of words required in making lexicons, we can refer to that fact that "current commercial MT systems typically claim to have general-language dictionaries containing upwards of 15,000 entries"[1].
3. The scope and quality of translation are to a great extent dependant on the lexicon. "More than any other component, the size and quality of the dictionary limits the scope and coverage of a system"[2].
4. A lexicon is the only component in MT that is always subject to expansion and updating. While adding or modifying morphological, semantic, or syntactic rules is not expected to be done frequently, a lexicon is always subject to addition and modification. This is due to the fact that morphological and syntactic rules of a language are stable, but the vocabulary of a language is constantly growing, as new words are borrowed or coined every day.

MT dictionaries are sometimes called lexical databases, because the items are stored in a database, and sometimes called lexicons, because of the different nature these dictionaries have from the conventional dictionaries. "The lexical information for MT differs in may respects from that found in conventional dictionaries."[3] MT lexicons are different from conventional dictionaries in the way they are structured as well as in the type of information they hold. For example, while conventional dictionaries provide information on word pronunciation, origin, synonyms, antonyms, as well as explanatory examples of usage, none of this information is relevant in MT. Meanwhile, MT lexicons give explicit information about grammatical category, semantic features, selection restrictions, subcategorization features, and case frames.

---

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 54.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 87.
[3] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 78.

We now turn to the structure of an MT lexicon and the type of information it contains.

1. Information on grammatical categories (or parts of speech). The grammatical categories usually have the typical abbreviated forms: 'n' for noun, 'v' for verb, 'pron' for pronoun, 'det' for determiner, 'prep' for preposition, 'adj' for adjective, 'adv' for adverb, and 'conj' for conjunction. The MT module responsible for identifying the part of speech (POS) is usually referred to as "Part-of-Speech Tagger"[1].

2. Subcategorization information. It "indicates the syntactic environments that a word can occur in."[2] The lexicon must indicate, for example, whether a verb is transitive or intransitive, and a noun is masculine or feminine. Verbs can have as many as eight different subcategories as in the following list:

1.	sleep [I]: An intransitive verb that only needs a subject. e.g. The old man slept.
2.	build [Tn]: A transitive verb that needs a subject and object. e.g. He built a house.
3.	give [Dn,n] A ditransitive verb that takes a subject and two objects. e.g. He gave Jack a book.
4.	give [Dn,prep] A ditransitive verb that takes a subject and two objects, one introduced by a preposition. e.g. He gave a book to Jack.
5.	persuade [Cn,to] A complex transitive verb that takes a subject, object and infinitival clause introduced by *to*. e.g. He persuaded Jack to travel abroad.
6.	believe [T,fn] A transitive verb that takes a subject and an object in the form of a finite sentence introduced by *that*. e.g. He believed that the problem was solved.
7.	is [La] A linking verb that links an adjectival phrase to the subject. e.g. This young man is handsome.
8.	appear [Ln] A linking verb that links a noun phrase to the subject.

[1] Fred Popowich, et al., "Machine Translation of Closed Captions", *Machine Translation*, 15, 311-341, 2000, p. 322.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 92.

	e.g. Jack appeared a handsome young man.

Table 2. Possible subcategories of verbs[1]

3. Semantic information. Semantics can be defined as "the study of the ways in which individual words (lexical items) have meaning either in isolation or in the context of words, and the ways in which phrases and sentences express meaning."[2] The importance of semantics in MT research comes form the fact that the semantic analysis of a language explains "how the sentences of this language are understood, interpreted, and related to states, processes, and objects in the universe."[3] Words can be represented in terms of semantic features, which can constitute a logical hierarchy as in the following figure:

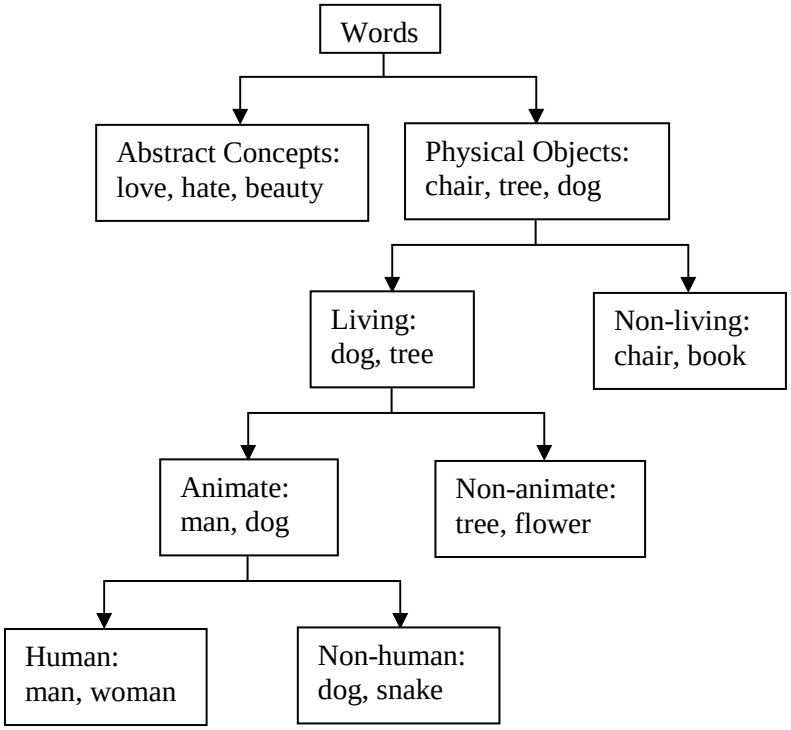


Fig. 8. The hierarchy of basic semantic features

Further semantic features, such as 'male' can be added to the above hierarchy to distinguish *man* and *boy* from *woman* and *girl*. The semantic

[1] Adapted from: Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 92.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 19.
[3] Manfred Bierwish, "Semantics", in John Lyons, ed., *New Horizons in Linguistics*, Middlesex, England: Penguin Books Ltd, 1970, p. 167.

feature ‘adult’ can furthermore distinguish *man* and *woman* from *boy* and *girl*. So the semantic features for the four words *man, woman, boy* and *girl* can be listed as follows:

1.	man: PHYSICAL OBJECT and LIVING and ANIMATE and HUMAN and MALE and ADULT
2.	woman: PHYSICAL OBJECT and LIVING and ANIMATE and HUMAN and FEMALE and ADULT
3.	boy: PHYSICAL OBJECT and LIVING and ANIMATE and HUMAN and MALE and YOUNG
4.	girl: PHYSICAL OBJECT and LIVING and ANIMATE and HUMAN and FEMALE and YOUNG

Table 3. List of semantic features of some words

This explains how the meaning of each word is composed of the set of semantic features attached to it. The above list of features can be further simplified by introducing the notion of ‘inheritance hierarchy’, “so that it is sufficient to say that the word *teacher* has the feature ‘human’ to know that it is also ‘animate’, ‘living’, and so on.”[1]

4. Selection restriction information. It is the restrictions a lexical item imposes on the words that occur with it. Semantic restrictions “indicate which lexical elements may be selected in order to form a semantically well-formed combination of two or more syntactically combined lexical elements. They specify … possible semantic affinities among lexical entries.”[2] For example the verb *read* requires a subject with the feature ‘human’ and the verb *drink* requires an object with the feature ‘liquid’.

All of the above information can be stored in a lexical database in the form of a set of features. Each feature “can be represented as **attributes** with corresponding **values**, and as such are often called ‘attribute-value pairs’.”[3] The following table gives a sample of the possible attribute-value pair in an MT dictionary:

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 19.
[2] Manfred Bierwish, “Semantics”, in John Lyons, ed., *New Horizons in Linguistics*, Middlesex, England: Penguin Books Ltd, 1970, p. 171.
[3] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 25.

man	CAT: n NUM: sg GEND: m SEM: human	(Grammatical category: noun) (Number: singular) (Gender: male) (Semantic feature: human)
drink	CAT: v SUBCAT: T PERS: NUM: SEM_AGENT: animate (Takes an animate subject) SEM_PATIENT: liquid (Takes a liquid object)	(Grammatical category: verb) (Subcategorization: Transitive verb) (Person)

Table 4. Feature analysis

To ensure consistency and completeness in building an MT lexicon, MT developers usually develop tools for lexicographers to aid them easily select the appropriate features for each lexical item. These tools provide linguists with "menus or templates for entering the lexical data which is required, e.g. the gender of a noun, whether it has an irregular plural, whether it is 'mass' or 'count'."[1] This type of programs also ensures that the dictionary will contain no duplicate entries. When the lexicographer wants to create a new entry, the program "searches the database for records already in existence."[2] If the entry already exists, the user is only permitted to modify its properties.

The design of an MT dictionary is to a great extent influenced by the translation strategy which a system follows. In a direct translation system, there is only one bilingual dictionary of source language words and their equivalents in the target language. By contrast, most transfer MT systems have two types of lexicons: monolingual lexicons for purposes of analysis and generation, and bilingual lexicons for mapping source language words with their equivalents in the target language.

The source language monolingual lexicon contains all information needed for structural analysis and disambiguation. It contains, for example, morphological inflections, grammatical categories, semantic features, selection restriction, different meanings of a word, and the conditions (whether syntactic, contextual, or stylistic) under which the word assumes a specific meaning. For example, the lexicon should be

[1] Ibid., p. 55.
[2] Laffling, John. *Towards High-Precision Machine Translation: Based on Contrastive Textology*. Berlin: Foris Publications. 1991 p. 65.

able to tell when the word *bank* means 'the side of a river' and when it means 'a financial institution'.

The bilingual lexicon will "list all target language equivalents"[1]. The bilingual lexicon may contain some grammatical and semantic information, yet this information is minimal, as it is properly detailed in the monolingual lexicons. The target language monolingual lexicon is responsible for the generation of target language words.

However, this does not mean that an MT system has only three dictionaries. In fact it can have several other dictionaries. MT lexicons can be divided into "a number of special dictionaries, e.g. for 'high frequency' vocabulary, idiomatic expressions, irregular forms, etc., which are separated from the main or 'core' lexicons."[2]

One of the major issues in the construction of MT lexicons is the treatment of terminology and idioms. Terminology is the vocabulary of a specialized subject field like biology, economics, and medicine. "Such fields often have a relatively well-defined terminology, which is sometimes even codified, and given official recognition by professional bodies."[3] Codification means that each concept is assigned a single name. This is why terminology is less ambiguous and can be more easily manipulated by an MT system than general language words. Whereas terminology dictionaries represent one-to-one mapping of words and concepts, words in a general language dictionary are ambiguous because a word can designate more than one concept. It is a regular tradition in MT systems to store terminology in separate specialized lexicons.

Idioms are "fixed phrases, consisting of more than one word, which have meanings which cannot be inferred by knowing the meanings of the individual words."[4] The following table lists some idioms along with their meanings:

---

[1] John Lehrberger and Laurent Bourbeau, *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation*, Amsterdam: John Benjamins Publishing Company, 1988, p. 57.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 79.
[3] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 107.
[4] Victoria Fromkin and Robert Rodman, *An Introduction to Language*, New York: Holt, Rinehart and Winston, Inc, 1974, p. 121.

	Idiom	Meaning
1	think the world of	have admiration for
2	pull strings	use influence
3	knock oneself out	work extremely hard
4	by the skin of one's teeth	just barely
5	upset the applecart	ruin plans
6	call a spade a spade	be completely frank
7	face the music	face a difficult situation
8	in a nutshell	in a few words
9	walk on air	be extremely happy
10	fly off the handle	suddenly get angry
11	burn the midnight oil	stay up late at night
12	fit to be tied	furious
13	pull oneself together	regain control of oneself

Table 5. List of some idioms[1]

From the above examples we see that the meaning of idioms is completely removed from the meaning of the single words that constitute them. Any MT system that does not make special consideration for idioms will give mistaken or even funny translations for the above idioms. "The problem with idioms, in an MT context, is that it is not usually possible to translate them using the normal rules."[2] So, Idioms must either have special entries in a lexicon or even a separate lexicon, and the system parser must treat them as a single unit.

Phrasal verbs pose a similar problem to an MT system. They cannot be translated as individual words but must be translated as a unit. For example, the meaning of the phrasal verb 'give up' cannot be conveyed by the translation of 'give' and 'up' separately. The two words combine together to give a new meaning. Another problem with phrasal verbs is that while in some phrasal verbs the particle cannot be split form the noun, such as 'look after', in some others the split can occur, as in the following examples:

He **gave up** smoking.
He **gave** smoking **up**.

[1] Idioms and their explanations are drawn from George P. MacCallum, *More Idiom Drills for Students of English as a Second Language*, New York: Harper & Row, Publishers, 1978.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 122.

He **looked up** the term.
He **looked** the term **up**.
He **filled in** the form.
He **filled** the form **in**.

What we need here is a way to relate the two parts of the phrasal verbs to each other. The MT parser must be furnished with all occurrence possibilities of each phrasal verb in order to be able to detect it whenever it occurs.

2.1.4. Syntactic Analysis

Syntactic analysis is the computer process of formalizing the "order and structure in a sentence and … the various relationships that words establish among themselves."[1] Syntactic analysis is one of the most problematic issues in MT and Computational Linguistics in general because it is related to the treatment of syntactic structure, which is far broader and more dynamic than the realms of morphology and lexicon. It poses a real challenge to linguists as well as computer engineers. Linguists have to search for the adequate theory that can formalize linguistic description; engineers have to find the adequate algorithms needed to process and analyze input. The process of syntactic analysis has to handle a large variety of sentence constructions and resolve many linguistic ambiguities. No single linguistic theory is said to have a paramount influence on MT design. Each MT system applies different theories and even devises some modifications or additional rules.

Whereas in morphological analysis the word is taken as the basic unit which is then analyzed into morphemes, in syntactic analysis the sentence is the basic unit which is then analyzed into constituents. It must be noted that sentences are not blocks similar to each other, but they are "surprisingly varied: long and short; simple, double, multiple, and complex; statements, commands, wishes, questions and exclamations"[2].

However, taking the sentence as a basic unit has two disadvantages. The first is that cross-sentential links (like pronouns referring to nouns mentioned in previous sentences) are not captured. The second is that it is difficult to decide exactly the sentence boundaries.

---

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 99.
[2] Simeon Potter, *Our Language*, Middlesex, England: Penguin, 1950, p. 90.

Whereas English sentences are known to begin with capital letters and end with full stops, sentences in other languages where punctuation does not play a vital role, like Arabic, cannot have any clear demarcating boundaries. An alterative definition of a sentence can be: "the largest linguistic unit to which syntactic rules apply."[1] This definition is somewhat circular in the sense that we need a definition of a syntactic rule in order to define a sentence, and we need a definition of a sentence in order to define a syntactic rule. In this thesis, however, I am only concentrating on translation from English into Arabic, so I will make use of the conventional definition of a sentence, that is: a sentence begins with a capital letter and ends with a full stop. The system, however, must not confuse a period that indicates an abbreviation such as Dr. and Mrs. and one that ends a sentence. The module responsible for splitting a paragraph into a group of sentences is called the 'segmenter'.

2.1.4.1. Phrase Structure Rules

Syntactic analysis starts with detecting the basic grammatical categories of V (Verb), N (Noun), Adj (Adjective), Adv (Adverb), Prep (Preposition), Det (Determiner) and Conj (Conjunction). There are four types of phrases:

1. VP: Verb Phrase
2. NP: Noun Phrase
3. AP: Adjective/Adverbial Phrase
4. PP: Prepositional Phrase

Analyzing a sentence into constituents in MT is called 'parsing'. Parsing means that the sentence is divided into smaller groupings according to their syntactic functions inside the sentence. After the sentence is parsed it is represented in a parse tree or phrase structure tree. The simple sentence *The young boy eats the apples* can be represented as follows:

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 13.

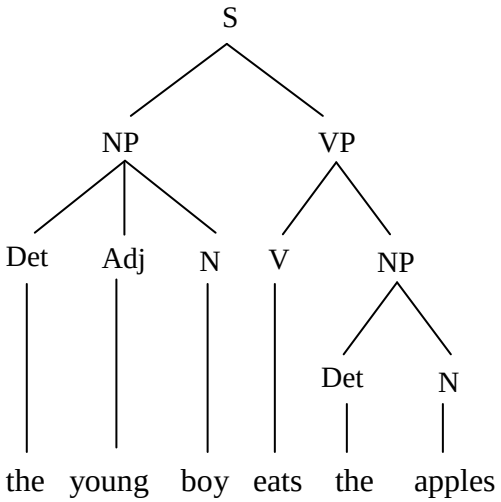


Fig. 9. A parse tree[1]

The above tree captures various grammatical relationships like dominance and precedence. Dominance means that some nodes dominate other nodes. In the above example S dominates NP and VP. Precedence means that some nodes precede other nodes. In the above example NP precedes VP. The mother node is the node that dominates other nodes. The first NP is the mother for Det, Adj and N, which in their turn are said to be daughters of the first NP and sisters of one another.

The above tree can be represented by re-write rules as follows:

(1) S → NP VP

(2) NP → (Det) (Adj) N

(3) VP → V NP

(4) Det → {*the*}

(5) Adj → {*young*}

(6) N → {*boy, apples*}

(7) V → {*eats*}

These re-write rules state that a Sentence (S) is composed of a Noun Phrase (NP) followed by a Verb Phrase (VP). The Noun Phrase (NP) consists of an optional Determiner (Det) and an optional Adjective (Adj)

[1] Adapted from Tim Willis, "Processing Natural Language", in Peter Roach, ed., *Computing in Linguistics and Phonetics: Introductory Readings*, San Diego: Academic Press, 1992, p. 60.

and a Noun (N). The Verb Phrase (VP) consists of a Verb (V) and a Noun Phrase (NP). The terminal nodes are re-written as lexical items.

An alternative representation to phrase structure tree is a bracketed string of categories and elements, which is even more suitable to an MT system. The same information in the Phrase Structure tree can be represented as follows:

```
S(NP(Det(the)
        Adj(young)
        N(boy))
  (VP(V(eats)
        NP(Det(the)
             N(Apples)))
```

There are three types of relationships that can be captured by syntactic description. The first is the relationship of 'sequence' which states which elements precede other elements. In the above tree, for example, the adjective *young* precedes the noun *boy*, which it modifies. The second type of relationship is 'dependence' which states that some elements are dependant on other elements. For example, the preposition can determine the morphological form (or case) of the noun that follows it, as in *to him*, in which case the pronoun is dependant on the preposition. In the above example, the noun *boy* and the verb *eats* both agree in number, an example of the verb dependence on the subject. The third type of relationship is **constituency** which means that some elements may be combined together to constitute a phrase. In the above example, the first NP consists of Det, Adj and N. These are some of the basic concepts needed in syntactic analysis.

2.1.4.2. Parsing

Parsing is a major component in an MT system. It is the process by which a sentence is dissected or analyzed into constituent parts. Parsing may be defined as the "computer programs which take as data a grammar and a lexicon … and produce as output an analysis of the structure of the text."[1] The MT module responsible for parsing is called 'the parser'. While the lexicon stores information about the categories of words and

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, pp. 56-57.

their idiosyncratic features, the grammar rules in the parser, according to Hutchins[1], fulfill the following two tasks:

1. Recognize what sequences of words form constituent phrases, which in turn combine to form bigger phrases or sentences.
2. Describe what constructions are acceptable (or grammatical).

Inside the parser, grammar rules are usually written as a set of re-write rules trying to give all possible variations of each constituent. We can have a second look at re-write rules in the following set of rules[2]:

(1) S → NP VP (PP)

(2) NP → Det (Adj*) N (PP)

(3) NP → Pron

(4) NP → NP S

(5) VP → V NP

(6) VP → V NP NP

(7) VP → V PP

(8) VP → V

(9) PP → P NP

The above re-write rules give some possible variations of S, NP and VP. An NP can be composed of a determiner, adjective, noun and prepositional phrase as in '*the young man with the blue hat*' (elements in parentheses in re-write rules are optional). The '*' beside 'Adj' in rule (2) means that the noun in a noun phrase can be modified by any number of adjectives. For example, we can add as many adjectives as we like before the word 'man'. We can say '*the handsome, tall, attractive, busy young man*'. There is no physical limit on the number of adjectives preceding a noun except the memory space of the speaker and his desire to be comprehensible to the hearers.

The noun phrase can also be a simple pronoun such as *he*. In (4) the noun phrase is composed of an NP and S as in '*the fact that the earth is round*'. We notice that the phrase label NP invokes itself in the right side of the rule. It also invokes S which in its turn can be composed of an NP.

[1] Ibid, p. 57.

[2] Adapted from Tim Willis, "Processing Natural Language", in Peter Roach, ed., *Computing in Linguistics and Phonetics: Introductory Readings*, San Diego: Academic Press, 1992, pp. 56-57.

We also notice that the NP and PP can both invoke each other. This explains the grammaticality of the children's song: '*There's a hole in the middle of the sea*' and its expansion: '*There is a bump on the flea on the lump on the fly on the hump on the frog on the stump on the log in the hole in the middle of the sea*'. This property is called 'recursion' which accounts for the infiniteness of the human language and the number of sentences it can contain. "It is a property that enables the small set of eventual rules … to describe the structure of, or generate, a very large number of the possible sentences"[1].

The verb phrase in turn can be composed of a verb and a noun phrase when the verb is transitive as in '*build a house*'; a verb and two noun phrases when the verb is ditransitive as in '*give John a book*'; a verb and prepositional phrase as in '*go to school*'; or only a verb when the verb is intransitive as in '*sleep*'. The information about the type of verb is gained from the lexicon.

This type of parsing is known as a 'top-down approach' where "the parser starts at the most abstract level and attempts to flesh out the structure by building downwards towards the lowest level, i.e. the words themselves."[2] Following this approach, the parser supposes that the string at hand is an S and searches for an NP followed by a VP. The NP requirement is fulfilled when the system finds any of the sequences that constitute an NP and so on.

In order for the rules to be comprehensive, they should cover other forms of the sentence. A sentence structure can vary according to mood, which can be either indicative, like '*He went to the school*'; subjunctive, like '*If he studied, he would succeed*'; imperative, like '*Open the door!*'; interrogative, like "*How are you?*'; exclamatory, like '*What a car!*'; or optative, like '*Had I the means, I would travel abroad*'. A sentence structure can also vary according to voice, which can be either active like '*I teach*' or passive '*I was taught*'. Moreover, sentences can be composed of more than one clause, like '*He is rich but he is not happy*'. It can also be composed of a non-tensed phrase plus a clause, like '*Going to school, I lost my bag*'. Moreover, some sentences may not contain a verb, like '*More work, more money.*'

[1] Ibid., p. 58.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 57.

As we can see in those examples, a sentence structure does not always follow the NP VP pattern. There is a variety of patterns that should be accounted for in the grammar if the MT system is to deal with the human language successfully, from the syntactic point of view.

Besides the top-down approach, there is the bottom-up approach where the parser "starts with the words and builds upwards."[1] The parser starts with recognizing word categories then searches for rules that allow for the combination of these words into higher levels of abstraction, or bigger phrases. Take for example the sentence '*He went to the garden*'. In parsing this sentence the parser, according to Hutchins[2], will follow these steps:

1. Recognizing word categories.

   Pron V Prep Det N

2. Parsing the first NP following rule 3 above, which states that an NP can be composed of a pronoun.

   NP V Prep Det N

3. Parsing the second NP following rule 2 above, which states that an NP can be composed of a determiner followed be a noun.

   NP V Prep NP

4. Parsing the PP following rule 7 above, which states that a PP is composed of a preposition and an NP.

   NP V PP

5. Parsing the VP following rule 5 above, which states that a VP can be composed of a verb and a PP.

   NP VP

6. Parsing the S following rule 1 above, which states that an S can be composed of an NP followed by a VP.

   S

The above six steps can be summarized in three rules for sentence parsing as follows:[3]

---

[1] Ibid.
[2] Ibid.
[3] Keith Brown and Jim Miller. *Syntax: A Linguistic Introduction to Sentence Structure*, 2nd ed., London: Harper Collins Academic, 1991, p. 46.

(1) Determine from the lexicon the form class of every word …
(2) Find a grammar rule with X as the left-hand symbol and a category or categories as the right-hand symbol(s) (for example, X → Y Z) and where the right-hand symbol(s) match some sequence of categories in the structure so far and develop a partial tree with X as the mother and the right-hand symbols as ordered daughters.
(3) Continue rule (2) until the tree is rooted in S and there are no unattached items.

Another concept of major importance in parsing is the notion of 'backtracking'. Backtracking means that the parser can make initial decisions about the category of a word, its meaning or its syntactic function in the sentence, and then makes revision of this initial decision when the context provides extra information as the parser proceeds with the sentence. "In both top-down and bottom-up parsing, it is not always the case that the first decision is the right one… the computer has to be able to revise any 'decisions' it may have taken."[1] Take for example the sentence:

The river flows harm riverside residents.

In a bottom-up approach, the parser can make the initial decisions that 'the river' is an NP and 'flows' is a V. But then it encounters the word 'harm'. If the system decides that 'harm' is a noun then the system will not succeed in designating a syntactic function for this noun as well as the following noun *riverside residents*. These nouns cannot be objects as the verb *flow* is not di-transitive. The best decision is that 'harm' is a verb. The system then has to revise its initial decisions about the starting part of the sentence. The block 'the river flows' is an NP, 'harm' is a V and 'riverside residents' is an NP.

In a top-down approach, the parser will make the supposition that the sentence has the NP VP form. Then it will look for the string that matches the NP requirements. It will make the initial decision that *the river* matches these requirements. Yet when it tries to search for a string that matches the VP requirements, it will fail. This failure prompts it to change the initial decision and take the string *the river flows* as an NP. The rest of the sentence will then match the specifications of a VP.

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 58.

2.1.4.3. Feature-Based Representation

Neither the phrase structure trees nor the bracketed strings discussed above are abstract or comprehensive enough to capture grammatical functions and represent syntactic and semantic features. A very good alternative is the feature-based representation which can represent all information needed in the form of attribute-value pairs. For example, 'category' is an attribute which can have 'noun', 'verb', 'preposition', 'adjective', 'adverb', or 'determiner' as a value. When the word is a noun the attribute 'gender' is relevant, and it can have either 'feminine', 'masculine' or 'neuter' as a valid value. The attribute 'number' is also relevant, and it can have either 'singular', 'dual' or 'plural' as a valid value. Similarly, when the word is a verb the attribute 'tense' is relevant, with either 'past', 'present' or 'future' as a valid value. Thus each word can have a feature or bundle of features. The set of features vary according to the word category. "Linguistic theories which use feature-based representations often include a **feature theory** which, for example, defines the lists of attributes and their possible values."[1] Using feature representation, our sample sentence '*The young boy eats the apples*' can be represented as follows[2]:

```
[cat:sentence
  dtr:{
        [cat:np, function:subj, num:sing,
           dtr:{
                 [cat:det, function:det, lex:the],
                 [cat:adj, function:mod, lex:young],
                 [cat:n, function:head, num:sing, sem:human, lex: boy]
                 }]
        [cat:vp, function:pred,
            dtr:{
                 [cat:v, function:head, tense:pres, lex:write, string:wrote],
                 [cat:np, function:subj, num:sing,
                    dtr:{
                          [cat:det, function:det, lex:the],
                          [cat:n, function:head, num:plur, sem:edible, lex: apples]
                          }]
                 }]
        }]
```

[1] Ibid., p. 25.

[2] Adapted from W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 26.

2.1.4.4. Parsing Problems

The process of parsing a sentence into constituent structures is a problematic issue. In many instances the category and meaning of a word (or even many words) in a sentence cannot be clearly defined. Some words may be assigned to more than one syntactic category or meaning according to the context. The parser has to apply different methodologies in order to resolve these lexical ambiguities. Another problematic issue that faces the parser is the syntactic ambiguities. It is not always clear which words form which constituents. Again the sentence may have different interpretations, and it is the task of the parser to resolve these ambiguities in order to arrive at a plausible interpretation of the sentence. In a human language it is "hard to find words that are not at least two ways ambiguous, and sentences which are (out of context) several ways ambiguous are the rule, not the exception."[1] In the next section I will explore these problems in more detail.

2.1.4.4.1. Lexical Ambiguity

The first problem which a parser meets when faced with a word is to define its grammatical category. Many words in English can be assigned to more than one category. For example the word *fast* can have one of four grammatical categories. It can be an adjective: '*This is a fast car*', an adverb: '*The car runs very fast*', a verb: '*Muslims fast during Ramadan*', and a noun: '*the fast of Ramadan*'. Another example is the word *light* which can also be assigned to the four categories. It can be a noun: '*the light of the sun*', a verb: '*They light the candle*', an adjective: '*a light bag*', and an adverb: '*He travels light*'. The ability to assume different grammatical categories is not limited to a small group of words. In fact it is a widespread phenomenon in English that "almost any noun can function as a verb."[2] This can be exemplified by the few examples given below:

[1] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 112.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 85.

Word	As a Noun	As a Verb
attack	The city came under attack during the night.	The enemy attacked us at night.
attempt	We failed in our attempt.	I attempted to leave but was stopped.
control	Which political party has control of the town council?	The pressure of steam in the engine is controlled by this button.
function	The function of a chairman is to lead and control meetings.	The machine won't function properly if you don't oil it.
heat	What is the heat of the water in the swimming pool?	We'll heat up some milk for the coffee.
help	If you want any help, just ask me.	Could you help me lift this box?
hint	There is a hint of summer in the air.	I hinted to him that I was dissatisfied with his work.
increase	Crime is on increase.	The population of this town has increased.
mistake	There were several spelling mistakes in your written work.	He'd mistaken the address, and gone to the wrong house.
paper	She papered the room with green paper.	She papered the room with green paper.
position	We can find our position by looking at this map.	He positioned himself by the entrance.
rise	a rise in the cost of living	The river is rising after the rains.
rule	It's against the rules to pick up the ball.	The king ruled for 30 years.
term	a medical term	The chairman of this parliament is termed the "speaker".
use	Do you approve of the use of guns by the police?	The company now uses a computer to do all its accounts.

Table 6. Examples of words functioning as nouns and verbs[1]

With some insight in the examples above, we can discern that the categorical ambiguity can be resolved either by inflection or textual context. When the word form acquires a morphological inflection that is specific for a certain category, then the ambiguity is solved. For example, when the word is suffixed by *-ed* or *-ing*, it becomes clear that the word in question is a verb. When this is not available, the parser has to find clues in the context to arrive at the category of the word in question. For example, when the word is preceded by an article (*a* or *the*) and/or followed by the preposition *of* then the only valid analysis is that the word is a noun. However, categorical ambiguity is not always so easily solved, especially when more than one ambiguity is found in the sentence. We have, for example, the sentence:

[1] Examples from *Longman Active Study Dictionary*, Essex: Longman, 1986.

Talk shows increase time waste.

Each word in the above sentence can be either a verb or a noun. No morphological inflection can resolve the ambiguity, since *shows* can be either a plural noun or a third-person singular present tense verb. No articles are found to give an indication of the category of any word. Only proper syntactic parsing of the sentence can resolve this ambiguity and give *increase* as the main verb of the sentence. Before arriving at this result the system can make several attempts and backtrackings. Statistically speaking, if the system tries to exhaust all possibilities, it will give two interpretations for each word and make the possible combinations with other words. In our example we have five words each with two possibilities: $2 \times 2 \times 2 \times 2 \times 2 = 2^5 = 32$. This means that the system will make 32 different interpretations of the sentence; all of them must be considered by the parser but only one of them is accepted as the correct one. This only gives a hint at the complexity of the issue.

Another type of lexical ambiguity is the problem of word meaning variations. Here a word, within the same part of speech, can have several different meanings. They are traditionally classified into homographs and polysemes. Homographs are words which happen to have the same spelling, but have quite different unrelated meanings and may be of different origins. Some well-known examples of homographs are *bank* as a river-side land or a financial body and *club* as a stick or a group of people. Homographs are usually given separate entries in dictionaries. Polysemes are words with different, but related, meanings, like the words *strike*, *game* and *tie* which have different related meanings as shown in the following list:

Homographs

No.	Word	Category	Meanings
1	bank	noun	1 The slope of land adjoining a body of water, especially adjoining a river, lake, or channel. 2 A business establishment in which money is kept for saving or commercial purposes. 3 A set of similar or matched things arranged in a row, especially: a. A set of elevators. b. A row of keys on a keyboard.
2	rail	noun	1 A steel bar used, usually in pairs, as a track for railroad cars or other wheeled vehicles. 2 Any of various marsh birds of the family Rallidae.
3	date	noun	1 Time stated in terms of the day, month, and year. 2 The sweet, edible, oblong or oval fruit of the date palm, containing a narrow, hard seed.
4	club	noun	1 A stout, heavy stick, usually thicker at one end, suitable for use as a weapon. 2 A group of people organized for a common purpose.
5	pile	noun	1 A quantity of objects stacked or thrown together in a heap. 2 A heavy beam of timber, concrete, or steel, driven into the earth as a foundation or support for a structure.

Polysemes

No.	Word	Category	Meanings
1	strike	noun	1 An attack, especially a military air attack on a single group of targets. 2 A cessation of work by employees in support of demands made on their employer, as for higher pay or improved conditions. 3 A sudden achievement or valuable discovery.
2	game	noun	1 An activity providing entertainment or amusement 2 A competitive activity or sport in which players contend with each other according to a set of rules. 3 Wild animals, birds, or fish hunted for food or sport. 4 An object of attack, ridicule, or pursuit.
3	tie	noun	1 A cord, string, or other means by which something is tied. 2 Something that connects or unites; a link: *a blood tie*. 3 A necktie.

Table 7. Examples of homographs and polysemes[1]

[1] Definitions from *The American Heritage Dictionary of the English Language*, 3rd ed., Boston: Houghton Mifflin Company, 1996.

The parser in an MT system does not have to make a distinction between homographs and polysemes. "In MT analysis homography and polysemy can often be treated alike, since it is a question of identifying the sense in context of a particular written 'word'."[1] The parser's problem here is to identify the intended meaning in a given context. We can now look at some sentences:

1. He went to the bank to cash a check.
2. He sat on the bank and looked at the clear water.
3. He walked beside the bank.

In order to solve the ambiguity of the word *bank* in the first sentence the system must make use of the semantic properties of other words in the sentence like *cash* and *check*, which state that they are related to money and finance. In the second sentence, the semantic restrictions of the verb *sit* require that it take as an object something that can be rested on. In this case a big building is excluded and a riverside bank is chosen. The word *water* mentioned also in the sentence gives emphasis to the choice. The third sentence is out of context and ambiguous both for a human reader and an MT system. Clues to solve such ambiguity can only be derived from the general context in which this sentence was said, or the subject field in which it was mentioned.

2.1.4.4.2. Syntactic Ambiguity

Syntactic ambiguity "arises when there is more than one way of analysing the underlying structure of a sentence according to the grammar used in the system."[2] Whereas lexical ambiguity involves possible different interpretations of a word, structural ambiguity involves possible different interpretations of the meaning of the whole sentence. Structural ambiguity, however, can be divided into two types: real ambiguity (when the sentence is ambiguous both for the human reader and the MT parser) and system ambiguity (when the sentence is ambiguous for an MT parser but clear to a human reader).

Language is abundant of real ambiguities, which, when taken out of context, can be confusing both for a human reader and a machine. Let us study, for example, the following sentence:

---

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 87.
[2] Ibid., p. 88.

The girl cleaned the table with a white cloth.

To a human reader this sentence can mean either the girl used a white cloth to clean the table or the table has a white cloth and the girl cleaned it. To the MT parser the PP, '*with a white cloth*' may be a constituent of VP following the re-write rule:

VP → V NP PP

This will generate the following parse tree:

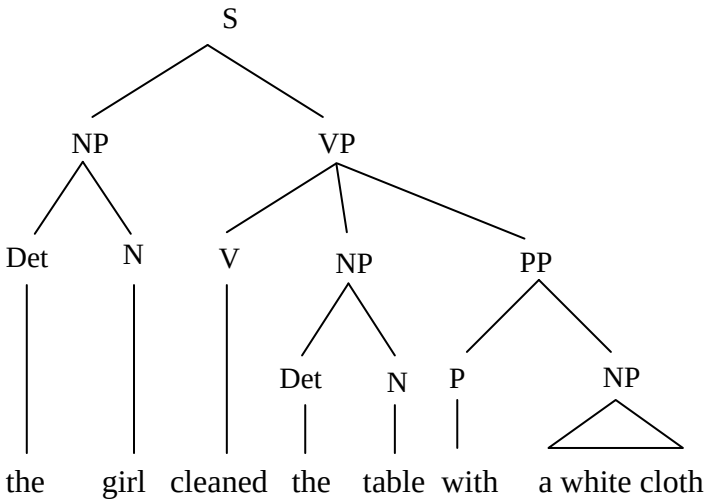


Fig. 10. First interpretation of an ambiguous sentence[1]

The other interpretation is that the PP '*with a white cloth*' can be a constituent of the NP, following the re-write rule:

NP → Det N PP

This will generate the following parse tree:

[1] Adapted from Tim Willis, "Processing Natural Language", in Peter Roach, ed., *Computing in Linguistics and Phonetics: Introductory Readings*, San Diego: Academic Press, 1992, p. 60.

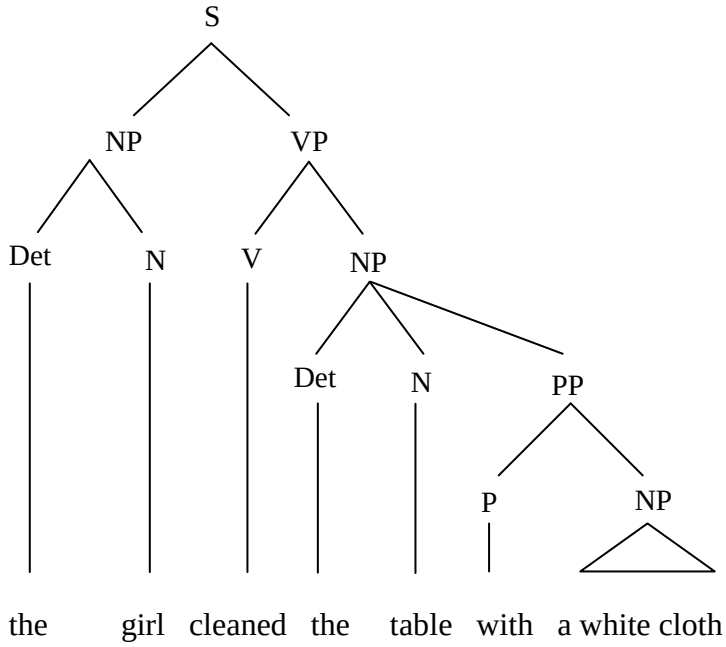


Fig. 11. Second interpretation of an ambiguous sentence[1]

This sentence gives no clues to solving its inherent ambiguity. In this case only the general "context will make the meaning clear, and sometimes it is simply impossible to resolve an ambiguity with certainty."[2]

Another example that illustrates structural ambiguity is the sentence introduced by Arnold et al.[3]:

> Cleaning fluids can be dangerous.

When a sentence has more than one interpretation, it is said to exhibit a deep structure ambiguity. Here the same surface structure has two different deep structures. This sentence has two possible interpretations, though the first interpretation is less plausible than the second:

> To clean fluids can be dangerous.
> Fluids which clean can be dangerous.

In the first interpretation '*cleaning*' is a gerund that governs a complement noun, and in this case '*cleaning*' is the head of the NP. In the second interpretation '*cleaning*' is an adjective modifying a noun, and in

[1] Adapted from Tim Willis, "Processing Natural Language", in Peter Roach, ed., *Computing in Linguistics and Phonetics: Introductory Readings*, San Diego: Academic Press, 1992, p. 60.
[2] Ibid., p. 59.
[3] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 113.

this case the noun '*fluids*' is the head of the NP. Deciding which word is the head of the NP affects meaning and also affects the structure if we use 'to be' instead of 'can':

Cleaning fluids is dangerous. (*cleaning* is head)
Cleaning fluids are dangerous. (*fluids* is head)

The other type of structural ambiguity is the system ambiguity. Sentences of this type are clear to a human reader but ambiguous to an MT parser. An example of this type is the prepositional-phrase attachment ambiguity, as shown in the following sentences:

He cut the book with a sharp knife.
He cut the book with a blue cover.

To a human reader, it is clear that '*a sharp knife*' is the instrument used in cutting. Therefore, the PP is an immediate constituent of the VP.

VP → V NP PP

In the second sentence '*a blue cover*' modifies the noun preceding it. The PP is an immediate constituent of the NP.

NP → Det N PP

To the machine, however, the two interpretations are possible with each of the two sentences. The parser does not have the kind of knowledge that makes it understand that a book cannot have '*a sharp knife*' as a modifier, and that '*a blue cover*' cannot be an instrument of cutting. If we want the machine to resolve this type of ambiguity and give a plausible interpretations of such sentences, ample "information should be added to the grammar so that the false reading is rejected."[1] There are different options for solving ambiguities in an MT model according to Hutchins[2]:

1. Using information from semantic features.
2. Using clues from the general context, i.e. clues found in previous sentences.
3. Using real world knowledge. This is a kind of knowledge added to the computer to let it know how things are in the outside world.

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 89.
[2] Ibid., p. 91.

4. Using interactive strategy; that is asking the system user to mark the correct choice.

Less plausible options are:

1. Using a default strategy; that is applying a generalization on all sorts of sentences. For example the computer might assume that all prepositional phrases are attached to the verb not the noun preceding it. This can make correct guesses in, say, 60% and false guesses in the other 40%.
2. Leaving the ambiguity as it is hoping that it will be preserved in the target language. This is what is termed 'free ride'. Free rides are sometimes used in MT when the target language has the same lexical or structural ambiguity. Here the system does not try to solve it, but transfers it at it is.

2.2. Transfer

Analysis is concerned with the representation of the source language by detecting constituent structures and resolving lexical and syntactic ambiguities. Generation is concerned with rendering the target language output in a grammatically acceptable form. Transfer is the interface or link between the two steps. Transfer "concerns the application of bilingual rules to the representations which result from analysis"[1]. In transfer all aspects of translational (whether lexical or structural) differences between the source and target languages are captured. Transfer starts with the output of the analysis phase, and ends where the phase of generation starts. There are three types of transfer discussed below: syntactic or structural transfer, lexical transfer, and morphological transfer.

2.2.1. Syntactic Transfer

Syntactic transfer provides the rules for converting source language parse trees into equivalent target language trees. It is a kind of "mappings

[1] Frank Van Eynde, "Machine Translation and Linguistic Motivation", in Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 73.

between the surface structure of sentences."[1] Lexical transfer takes care of all structural differences between languages regarding:

1. The distribution of subject-verb-object order (SVO) in a sentence. While English has the structure of SVO, Arabic has the preferred structure of VSO. Moreover, Arabic can transpose the subject and the object VOS, and can also front-shift the subject SVO.
2. The interpretation of verb tenses. While English has a variety of tenses to express aspects (whether the action is complete, progressive, or habitual), Arabic has only three tenses: present, past, and future. To map other English tenses into Arabic requires the use of temporal adverbials or helping words.
3. The position of adjectives and nouns (whether the adjective precedes or follows the noun).
4. The determiner system. English determiners include *a*, *the*, *this*, *these*, *that*, and *those*. The Arabic determiner system is different. It does not have an indefinite article, and (with demonstratives) it makes further distinction based on the gender of the modified noun.
5. The formation of passive sentences. Usually the passive has a different structure than that of the active voice.
6. Modal verbs used to express necessity, obligation, ability, intention, desire, and so on. Arabic does not necessarily use the same system for expressing the same functions.
7. Negation. Languages usually use different structure in negative sentences as contrasted by affirmative ones.
8. Mood. Lexical structures differ according to mood, whether it is declarative, imperative, interrogative, or exclamatory. Detailed mapping of different moods is needed in the transfer between English and Arabic as the two languages exhibit considerable difference in the way they express any of these moods.

Sometimes the mapping between source and target sentences requires little or no change in the target structure, such as the following two sentences:

The boy eats the apples.
ʾal-waladu yaʾkulu-t-tuffāḥ.

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 122.

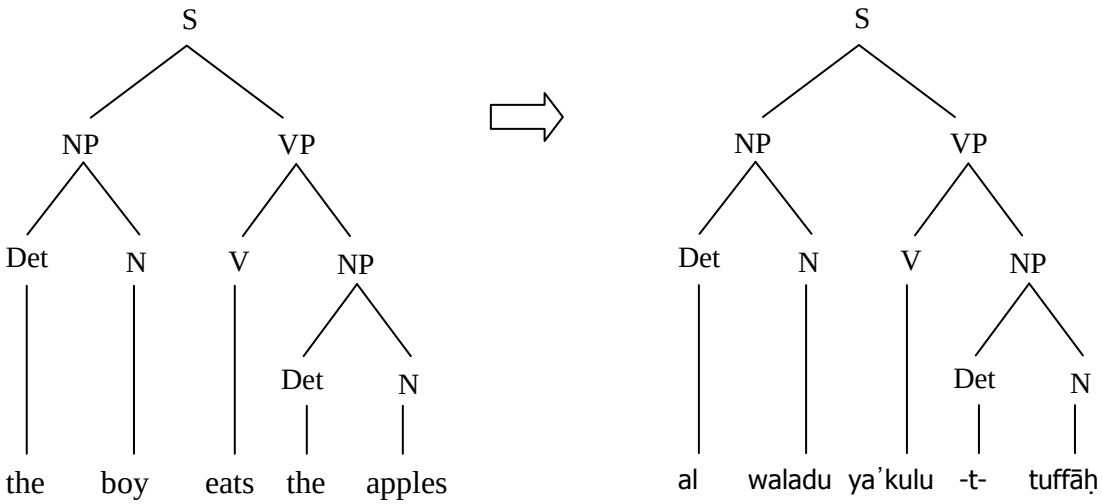


Fig. 12. Syntactic transfer

We notice that the output phrase structure is identical to the English structure. However, other changes to the sentence can be made during the phase of generation to ensure that the Arabic sentence bears the attributes unique to the Arabic language. The only recommended change here is to make the verb precede the subject. This permutation is to be made in the generation phase.

The reason we delayed adjusting the Arabic word order of VSO to the generation phase is that this word order is analyzed as a kind of movement or 'raising' in the surface structure and not inherent in the deep structure of the sentence. Within the framework of X-bar theory, the VSO order "is driven at S-structure by raising V to I. The subject may also raise to Spec of IP [Inflectional Phrase], which results in an SVO structure."[1] This movement can be illustrated in the following tree.

[1] Abdelkader Fassi Fehri, *Issues in the Structure of Arabic Clauses and Words*, Dordrecht: Kluwer Academic Publishers, 1993, p. 16.

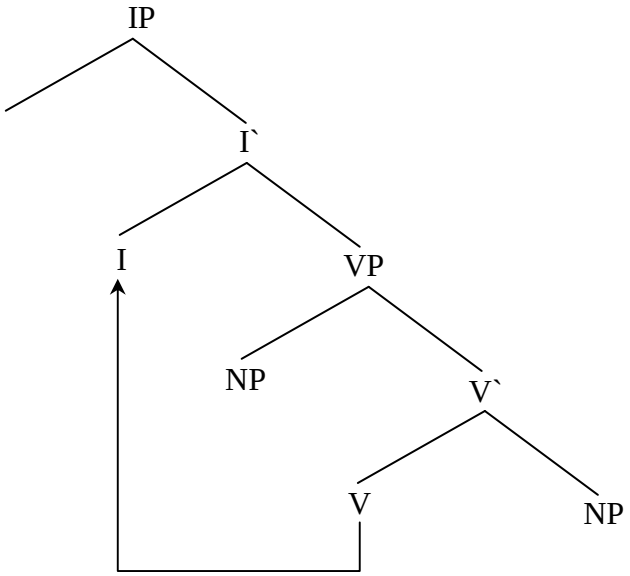


Fig. 13. V-to-I movement[1]

With a simple addition to the above sentence, we can make it a little more complex. Complex sentences require structural changes in the target language. Let us look at the following example:

The young boy eats the red apples.
yaʾkulu-l-waladu-ṣ-ṣaġīru-t-tuffāḥa-l-aḥmar.

What we need here is a rule for transferring English NP into proper Arabic NP to reflect the position of the adjective to the noun:

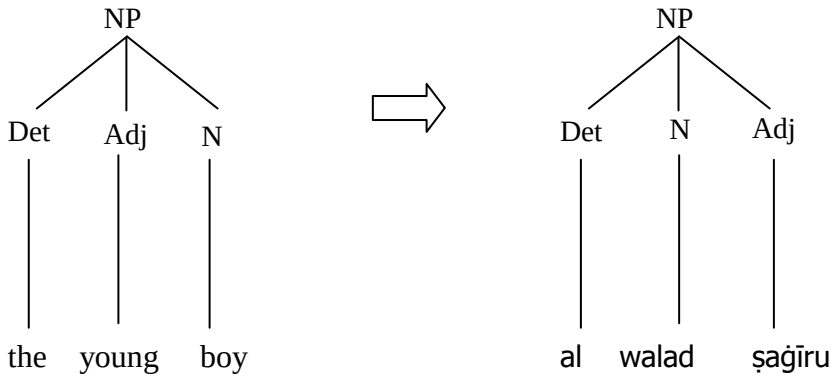


Fig. 14. Syntactic transfer of an NP

The transfer rules apply to each sub-tree of the structure until all constituents are appropriately dealt with. A "collection of tree-to-tree transformation is applied recursively to the analysis tree of the SL

[1] Ibid.

sentence in order to construct a TL analysis tree."[1] The tree-to-tree transformation algorithm is a recursive, top-down process in which rules are applied to all constituent structures until there are no other untransformed phrases. The transfer for the whole sentence will then look like this:

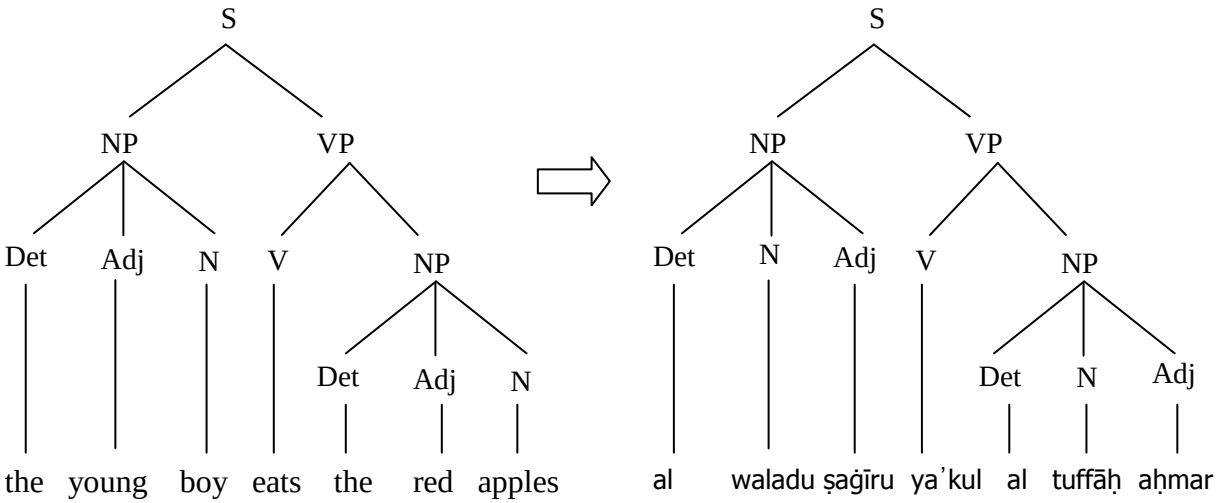


Fig. 15. Syntactic transfer

We notice that the output phrase structure still bears much similarity to the English structure. Some work is still required in the Arabic NP, that is the definite article which precedes the noun must also precede its adjective. We need also to swap the verb and subject to arrive at the Arabic preferred sentence structure of VSO. This kind of permutation, however, can be taken care of during the phase of generation.

Another instance that shows how the transfer rules should deal with differences between English and Arabic NP's is the process of nominalization, such as the following example:

English: NP → Det N PP
Example: the history of the world
Arabic: NP → N NP
Example: tarīḫu-l-ʿālam

Here the Arabic translation of the English NP has a different constituent structure. However, if the first noun is modified by an adjective the equivalent structure in Arabic will again be different from the first example.

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 122.

English: NP → Det Adj N PP
Example: the modern history of the world
Arabic: NP → Det N Det Adj PP
Example: ʾat-tārīẖu-l-ḥadīṯu li-l-ʿālam

There are other instances when the Arabic tree will be further distanced from the English tree. For example when the English sentence contains the verb *is* as a copula.

The flower is beautiful.
ʾaz-zahratu ğamīlah.

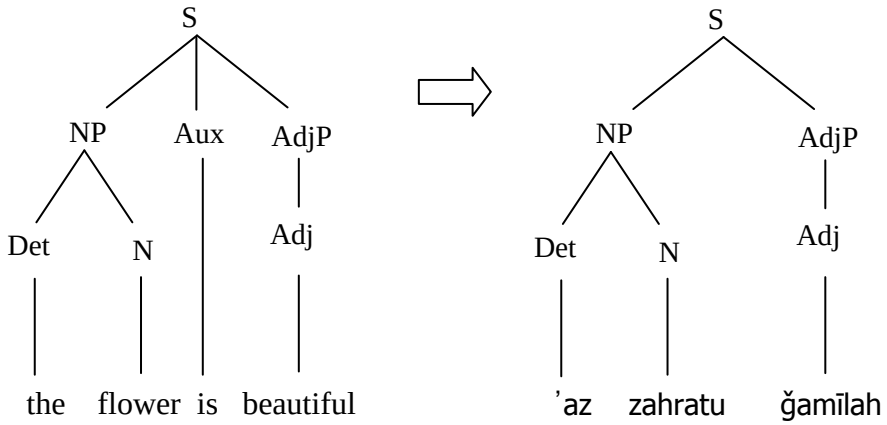


Fig. 16. Syntactic transfer

However, if we use the past form of the verb *was* or *were*, a verb (*kāna*) will be needed in Arabic.

Another type of sentences that exhibit differences between English and Arabic structures is the passive. The English passive is composed of *be* and the past participle form of the verb, and the object is moved to the position of the subject. In contrast, the Arabic passive is made by a different morphological form of the verb and alternative case marking on the object.

The boy was hit.
ḍuriba-l-waladu.

The formation of the negative in English exhibits, as well, structural differences between English and Arabic. English negative sentences can have different forms:

1. Using *not* after the copula *be*:

The boy is not happy.

2. Using *not* after the auxiliary:

The boy has not gone to school.

3. Using a form of *do* and *not* and changing the main verb into the infinitive:

    He did not play.

4. Using *no* before the noun:

    The boy gave no attention to the idea.

5. Using *never*:

    He never wakes up early.

Some of the Arabic rules for negation are:

1. Using *laisa* in subject-predicate sentences and changing the case marking for the predicate:

    Laisa-l-waladu saʿīdan.

2. Using *lā* with present tense sentences:

    lā yaḏhabu-l-waladu ʾilā-l-madrasati

3. Using *lan* with future tense sentences:

    lan yaḏhaba-l-waladu ʾilā-l-madrasati.

4. Using *mā* with past tense sentences:

    mā ḏahaba-l-waladu ʾilā-l-madrasati.

Mapping English negative structures into the Arabic negative structures is the job of detailed and exhaustive transfer rules. Another possible alternative is only to mark the sentence as 'negative' and leave the transformation for the generation phase which will take care of the necessary changes.

2.2.2. Lexical Transfer

Lexical transfer means "the replacement of a source lexical item by a target lexical item."[1] The hypothetical ideal world for the transfer of lexical items from a source language into a target language is that each source word has only one equivalent target word. This can be the case only in technical documents where unified terminology is used. A unified terminology, which uses one-to-one correspondences, serves both in consistency and clearance of concepts. This is why MT proved very

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 113.

useful and successful in the translation of technical documents. An example of this is the terminology of computer terms.

English	**Arabic**
printer	ṭābiʿah
monitor	šāšah
compact disk	qurṣ mudmağ

Table 8. Transfer of technical terms

In everyday language, however, this is seldom the case. We are met with words which can assume a variety of meanings as I have said in discussing the problems of homographs and polysemes. After deciding the intended meaning in the context the system must look the word up in the multilingual lexical database to get the proper translation of the word.

Sometimes an MT system meets a lexical transfer problem whenever faced by a translational ambiguity. This type of ambiguity occurs when a source language word can be translated by two or more target language words. The source word itself is not ambiguous, but the ambiguity emanates from the fact that the target language breaks down the concept into different concepts. This represents the one-to-many translation problem as exemplified by the following instances:

English	Arabic
dates	balaḥ (fresh) ruṭab (ripe) tamr (dry)
old	ʿağūz (with animate objects) qadīm (with inanimate objects)

Table 9. One-to-many lexical transfer

We notice here that the variety of translations for each word is not because each source word is polysemous, but because of conceptual translational ambiguity. This problem arises "when a single 'concept' represented by one word in one language corresponds to a number of concepts, and hence words, in another language."[1] The English culture has one concept for dates while the Arabic culture conceives them in many different ways. These conceptual differences are "reflections of

[1] Ibid., p. 101.

environmental or cultural differences"[1] between societies. The general trend in human and machine translation is to translate the word by a default equivalent in the target language unless the context provides clues that indicate otherwise.

By contrast, there may be a variety of concepts in the source language represented by different words; yet they are regarded as one concept in the target language. This represents the many-to-one case. This is not, however, a problem in translation as they are all translated by the same word.

English	**Arabic**
hire / rent	yastaʾǧir
pigeon / dove	ḥamāmah
library / bookshop	maktabah
snow / sleet / slush / hail / blizzard	ǧalīd

Table 10. Many-to-one lexical transfer

The above examples show some conceptual differences between English and Arabic. In English there are different concepts for the temporary use of service or property for a fee (hire) or for some time for a regular payment (rent). From the perspective of Arabic, these two concepts are regarded the same. Another example which represents the environmental differences between the Arabic speaking society and the English one is represented by the variety of English words for snow.

2.2.3. Morphological Transfer

In an MT transfer process mapping is not made between inflected lexical items because this will mean that monolingual morphological components of the source and target languages are incorporated in the transfer component. This, of course, will be a big burden beyond the capacity of the transfer module. This will also violate the modularity principle in programming. This principle states that a big process, such as translation, be divided among different, separate -- and at the same time integrated -- components. In an ideal MT system, transfer is carried out at the uninflected lexical items. In our tree, all inflections expressing grammatical functions are removed and expressed as feature notations.

[1] Ibid., p. 102.

For the sentence '*The boy eats the apples*', the complete transfer structure will look like this:

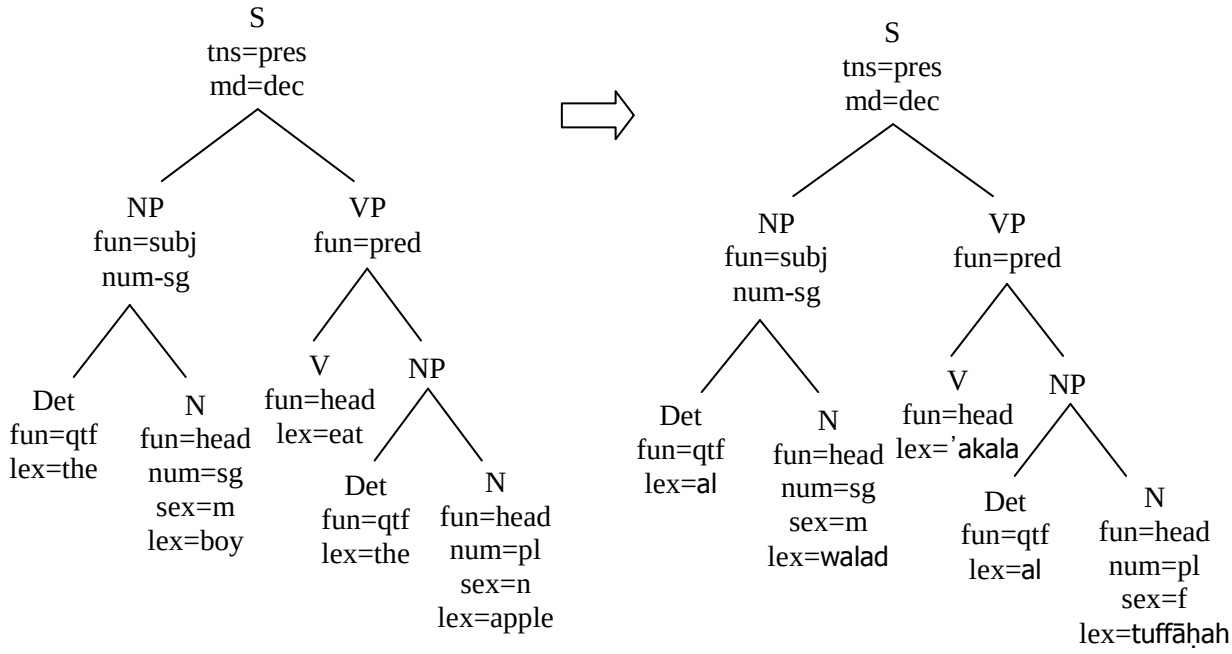


Fig. 17. Morphological transfer

Some abbreviations are used in this model for purposes of simplicity and brevity. Please see the list of abbreviations at the beginning of this thesis.

All words in the above transfer model are rendered in their uninflected forms. Nouns are all in the singular and verbs in the infinitive. Number and gender markings of nouns are given as features of each noun, and the tense is given as a feature of the whole sentence. "Grammatical information such as tense and number has … been passed up to the highest appropriate node."[1] The gender of a noun may differ from one language to the other. As we notice in the example the English *apple* is neuter while the Arabic *tuffāḥah* is feminine. The analysis here is deeper than in the previous models and so the transfer is simpler as it does not have to worry about morphological problems. Transfer rules are primarily concerned with structural changes between the two sentences. The generation component will look after the transformation of the output structure into a grammatically and morphologically acceptable target sentence.

[1] Ibid., p. 111.

2.3. Generation of the Target Language

The last step in the translation process is the generation of the target language input. Generation "is the process of deriving linguistic expressions whose meanings correspond to some informational structure used as input."[1] It takes as input the structures and lexical items produced from the transfer phase, and then produces well-formed, grammatically accepted target language sentences. The generation process "involves only target language information and operates independently of the source text."[2]

Generation was considered by earlier linguists and MT engineers as a simple, straight-forward process. They did not give this process the same attention as that given to either analysis or transfer. Therefore, the output mostly had a rigid and mechanical style. Yet this view has changed. As MT translation is judged by the quality of its output, researchers began to underscore the importance of generation. It is now understood that a "good MT system must be able to appreciate the nuances of the word choices in the target language (TL) and the interfaces that are invited by alternative syntactic phrasings."[3] A low quality translation is one which bears a lot of the syntactic and lexical characteristics of the source language. This is what translators used to label as 'literal translation'. Therefore, the quality of input can be greatly fine-tuned by making the generation component of an MT system produce sentences that carry little characteristics of the source language.

Generation takes the representation produced by transfer as input and produces the target language sentences. This process is usually split into two components: syntactic generation and morphological generation.

2.3.1. Syntactic Generation

The input from transfer is a deep-structure representation of the target language. The tree is labeled with features that indicate whether the

[1] Arturo Trujillo, *Translation Engines: Techniques for Machine Translation*, London: Springer, 1999, p. 111.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 132.
[3] David D. McDonald, "Natural language generation: complexities and techniques", in Sergei Nirenburg, ed., *Machine Translation: Theoretical and methodological issues*, Cambridge: Cambridge University Press, 1987, pp. 22-23.

sentence is active or passive, negative or affirmative, indicative, interrogative, imperative or exclamatory. The syntactic generation component uses transformational rules to convert this deep structure "into an ordered surface-structure tree, with appropriate labelling of the leaves with target language grammatical functions and features. The basic task of syntactic generation is to order constituents in the correct sequence for the target language."[1]

For example, in our English-Arabic context, if the sentence is labeled 'interrogative' in the deep structure representation, the generation component will choose the appropriate question word in Arabic (whether *man, māḏā, kaifa, lima, hal,* etc.) and rearrange the constituents according to the Arabic word order in interrogative sentences. If a sentence is marked 'passive' the generation component will label the verb with the feature 'passive' and assign the 'nominative case' feature to the object. The realization of these features by choosing the right morphological forms of the words will come later in the morphological generation phase.

The syntactic generation component will carry out all changes required by the target lexical items. As words project onto the syntax, they require a different environment from that of the source sentence. For example, compare the following English sentence and its Arabic translation:

> The student keeps the book.
> aṭ-ṭalibu yaḥtafiẓu bi-l-kitāb.

We notice here that while the English verb 'keep' takes a direct object, the Arabic verb '*yaḥtafiẓu*' takes a prepositional phrase. The syntactic module will handle this modification and assign the appropriate case to the prepositional object.

Another task of the syntactic generation component is "the distribution of number and gender information to relevant terminal nodes."[2] It must ensure in Arabic that the adjectives agree in number and gender with the noun they modify, verbs agree with subject nouns, pronouns agree with the nouns they refer to, demonstratives and relative pronouns also agree with the referent nouns. When information needed for agreement is not

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 133.
[2] Ibid., p. 134.

provided during analysis and transfer, the generation component must adequately supply this information.

2.3.2. Morphological Generation

The morphological generation module "interprets strings of labeled lexical items for output as target sentences"[1] After all constituents are assigned appropriate features, the morphological module combines these features to produce the suitable inflected form of the words. For example, the Arabic verb *akala* (to eat) may be assigned the features [passive + past + 3rd person + feminine + single]. The verb combined with all these features will be rendered as *ʾukilat*. Similarly, the verb *ḏahaba* (to go) with the features [indicative + present + 3rd person + male + plural] will be rendered as *yaḏhabūn*.

It must be noted that the morphological generation component must be sensitive to word case. Arabic has a relative freedom in deciding the position of subject and object; it can have the order VSO or VOS. The lexical function is known by case marking. The following two sentences have more or less the same meaning:

ḍaraba-l-mudarrisu-ṭ-ṭāliba.
hit the teacher (NOM) the student (ACC)
The teacher hit the student.

ḍaraba-ṭ-ṭāliba-l-mudarrisu.
hit the student (ACC) the teacher (NOM)
The teacher hit the student.

Case marking is the primary indicator of the function of each constituent. Case is realized as a word ending or certain inflection of the word.

Besides case the system must be sensitive to irregular forms. The regular rule in Arabic masculine plural nouns is the suffixation of the morpheme *-ūn* (sometime *-īn*, according to case). Yet, there is a large number of irregular forms, as contrasted in the following examples:

[mudarris + plural + nominative] = mudarrisūna
[mudarris + plural + accusative] = mudarrisīna

[ṭālib + plural + nominative] = ṭullābun
[ṭālib + plural + accusative] = ṭullāban

[1] Ibid., p. 133.

The only thing that distinguishes nominative and accusative forms of the irregular plural noun *ṭullāb* is a diacritic case marking at the end of the word, usually ignored in modern writing.

The following figure represents the basic processes and components of an MT system. It summarizes all steps which a source languages text goes through until the equivalent target text is produced.

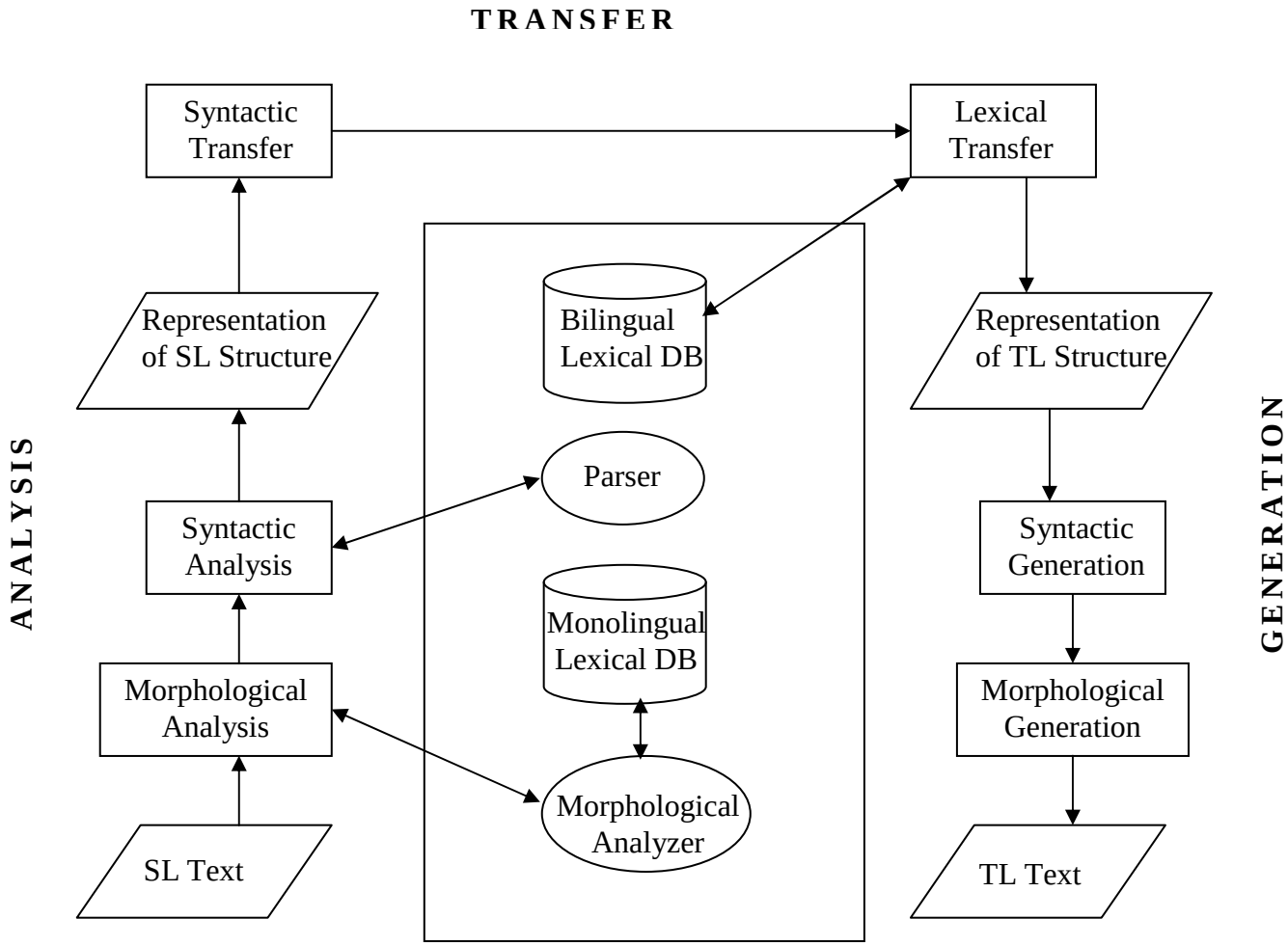


Fig. 18. Configuration of a transfer MT system

Chapter Three
The Agreement Feature

3.1. Definition

Agreement (sometimes referred to as concord) is defined as "a relation between words that share a morphosyntactic feature."[1] A more elaborate definition is the "correspondence in gender, number, case or person between words."[2] It is a grammatical relationship between two elements, which requires one element (the target) to have the same specifications for gender, number, case, person, or definiteness as that of another element (the controller). A target is said to agree with a controller when all or some of the agreement specifications are realized. The controller is defined as "the referent or head noun which serves to control the feature specification of the agreeing term or target."[3] The target is the word that carries these features.

The agreement relationship is seen by some linguists as a kind of syntactic linkage between words or phrases. When two constituents are in agreement with each other this means that there is a "tie that cuts across hierarchical structure"[4] to link these elements together. No matter how far the constituents from each other, if this special linkage exists, they must share the specified features.

The agreement relation between elements is seen by Koutsoudas as a special kind of co-occurrence relation. When we say that element X agrees with element Y, this means that "whichever grammatical categories are a property of Y are also a property of X whenever X and Y occur in the same well-specifiable positions."[5] This definition, however, is not precise because the claim that agreement requires that all grammatical categories which are a property of Y are also a property of X is not consistent with reality. Even in the example given by Koutsoudas,

---

[1] P. H. Matthews, *Syntax*, Cambridge: Cambridge University Press, 1981, p. 246.
[2] Definition from *The American Heritage Dictionary of the English Language*, 3rd ed., Houghton Mifflin Company, 1996.
[3] Reima Al-Jarf, "Grammatical agreement errors in L1/L2 translations. IRAL: *International Review of Applied Linguistics in Language Teaching*, 38.1, 2000, 1-15.
[4] Charles F. Hockett, *A Course in Modern Linguistics*, New York: The Macmillan Company,1958, p. 214.
[5] Andreas Koutsoudas, *Writing Transformational Grammars: An Introduction*, New York: McGraw-Hill, 1966, p. 131.

"*the girl runs quickly*," we see that the verb shares the grammatical categories of person and number with the subject, but not gender. A better definition would be to say that: an element is in agreement relationship with another element when one or more grammatical categories are shared by the two elements when they "occur in some well-specifiable positions."

Agreement rules are "rigid and well defined, and easy to code in a feature-based formalism."[1] In order to implement this formalism, "a feature called `agr`, whose value is a bundle representing the category's person, number, and gender, is added to the relevant categories."[2] For example, the agreement feature notation for the sentence:

He speaks English.

can be represented as follows:

(he [agr=x, number=sg, person=3rd])
(speak [agr=x, number=sg, person=3rd]) (English)

However, despite the fact that agreement rules are rigid and straightforward, many agreement problems arise during translation. This is due to the fact that agreement plays considerably larger roles in some languages than others.

Agreement as a unique grammatical relationship must be differentiated from both **government** and **dependency** relations. Some linguists may think that there is a relation between agreement and government and dependency, such as Hutchins who states that, "the sources of agreement relations lie in the governor in a dependency relation or in the head of a phrase structure."[3] The government theory states that each constituent within the sentence is composed of a head element and complements. The head is the noun in an NP, the preposition in a PP, the Verb in a VP, etc. In a dependency theory relations in a sentence are defined in terms of governor and dependents as in the following model:

The young boy likes the red apples.

---

[1] Manny Rayner, et al., eds., *The Spoken Language Translator,* Cambridge: Cambridge University Press, 2000, p. 16.
[2] Ibid., p.178.
[3] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 17.

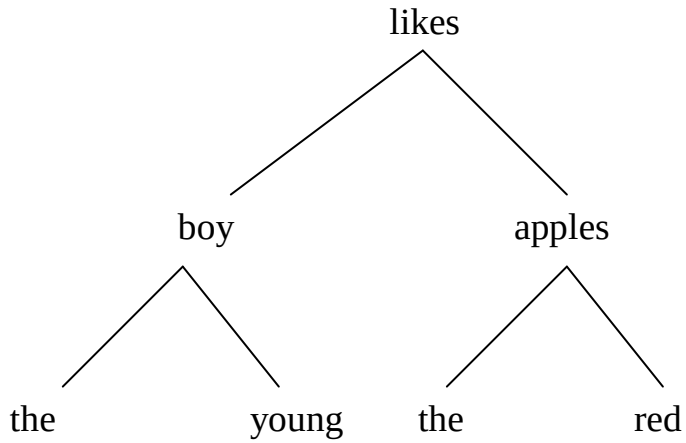


Fig. 19. Dependency tree

Agreement, however, is distinguished form government in that "the term 'government' is traditional only when the bound term marks a complement of the word which selects it."[1] Agreement, by contrast, displays a relationship beyond that of the head and complement. Regarding dependency, "it appears that agreement follows the direction of dependency when the dependent is a modifier or a determiner… but is the opposite"[2] elsewhere. In the above tree, whereas the verb is the governor and the subject is governed in the dependency model, it is the other way round in the agreement model.

A controller can be either a head noun or a noun phrase, while a target can be an article, adjective, verb, demonstrative pronoun, relative pronoun, or anaphoric pronoun. An agreement error occurs when there is "a mismatch in number, gender, case, or person between the subject and verb, adjective and modified noun, demonstrative and head noun, anaphoric or relative pronoun and antecedent."[3] Therefore, we can list the types of agreement as in the following section.

[1] P. H. Matthews, *Syntax*, Cambridge: Cambridge University Press, 1981, p. 248.
[2] Ibid., p. 250.
[3] Reima Al-Jarf, "Grammatical agreement errors in L1/L2 translations. IRAL: *International Review of Applied Linguistics in Language Teaching*, 38.1, 2000, 1-15.

3.2. Types of Agreement

In this section I will investigate the constituents (targets or controllees) that display agreement relationship with other constituents (controllers). I will draw on examples from English, Arabic and French. Examples from French are meant to bring more contrasts in order to shed more light on the effect of agreement on the morphology and syntax of sentences. French morphology is far richer than that of English and so agreement behavior in French is more perceptible.

3.2.1. Article-Noun Agreement

An example for article-noun agreement can be found in French where definite and indefinite articles are marked both for gender and number and are strictly required to agree with the noun they modify.

le garçon	la fille
the boy	the girl
les garçons	les filles
the boys	the girls
un garçon	une fille
a boy	a girl
des garçons	des filles
boys	girls

Here the gender and number of the controller (garçon/fille), which is the head noun, affects the choice of the article. Yet we notice that when the noun is plural the definite and indefinite articles are marked for number but not for gender. Moreover, the French number system does not include dual.

In English, the indefinite articles agree with the nouns they modify in number. If the nouns are singular, they are preceded by the article *a*, but if they are plural, they are preceded by an empty article, Ø.

a boy
boys

This type of agreement, however, disappears in the definite articles *the* which is not marked for number and precedes both singular and plural nouns.

the boy
the boys

This type has no example in Arabic, which has only one unmarked definite article, *ʾal-*, and no indefinite article at all.

walad	ʾal-walad
a boy	the boy
ʾawlād	ʾal-ʾawlād
boys	the boys

3.2.2. Adğective-Noun Agreement

This type of agreement is not found in English. Arabic, however, requires that the adjectives agree in number, gender, case and definiteness with the head nouns.

aṭ-ṭālibu-l-muğtahidu
the diligent (m, sg) the student (m, sg)

aṭ-ṭālibatu-l-muğtahidatu
the diligent (f, sg) the student (f, sg)

aṭ-ṭālibāni-l-muğtahidāni
the diligent (m, dl) the students (m, dl)

aṭ-ṭālibatāni-l-muğtahidatāni
the diligent (f, dl) the students (f, dl)

aṭ-ṭullābu-l-muğtahidūna
the diligent (m, pl) the students (m, pl)

aṭ-ṭālibātu-l-muğtahidātu
the diligent (f, pl) the students (f, pl)

In agreement, definiteness has a role to play: it is one of the features of agreement between adjectives and the nouns they modify. "In Arabic, a noun may or may not have an article… But if the noun has one so must an attributive adjective."[1] This is termed "agreement in definiteness"[2] and can be shown by the following examples:

baitun ṣaġīrun
house small
a small house

ʾal-baitu-ṣ-ṣaġīru
the house the small
the small house

[1] P. H. Matthews, *Syntax*, Cambridge: Cambridge University Press, 1981, p. 147.
[2] Joseph H. Greenberg, "Gender Markers", in Joseph H. Greenberg, ed., *Universals of Human Language*. Stanford, California: Stanford University Press, 1978, p. 50.

French also requires that the adjective agree in number and gender with the head noun.

le petit garçon the young boy	la petite fille the young girl
les petits garçons the young boys	les petites filles the young girls

We notice, however, that the agreement system in French adjectives is not as complicated as that in Arabic, since French has no case or definiteness agreement.

3.2.3. Verb-Subǧect Agreement

This type of agreement is found in many languages, yet the degree of agreement varies considerably. The English verb agrees with the subject only in person and number. When the subject is a third person singular and the tense is present, the verb is marked by suffixing *-s* or *-es*. In any other case the verb is not marked.

French shows a more complex system in verb agreement. The verb has to agree with the subject in number and person.

je vais. I go.	Nous alons. We go.
Tu va. You go.	Vous alez. You (pl) go.
Il vien. He goes.	Ils vonts. They (m) go.
Elle vien. She goes.	Elles vonts. They (f) go.

Arabic shows yet a more complex system in verb agreement than French as the verb agrees with the subject in person, number, and gender.

ʾanā aḏhabu.
I go (m/f).

naḥnu naḏhabu.
We go (m/f).

ʾanta taḏhabu. You (m, sg) go (m, sg).	ʾanti taḏhabīna. You (f, sg) go (f, sg).

ʾantumā taḏhabāni.
You (m/f, dl) go (m/f, dl).

ʾantum taḏhabūna. You (m, pl) go (m, pl).	ʾantunna taḏhabna. You (f, pl) go (f, pl).
huwa yaḏhabu. He goes.	hiya taḏhabu. She goes.
humā yaḏhabāni. They (m, dl) go (m, dl).	humā taḏhabāni. They (m, dl) go (m, dl).
hum yaḏhabūna. They (m, pl) go (m, pl).	hunna yaḏhabna. They (f, pl) go (f, pl).

Here we have 13 different forms of the verb in the present tense to express the agreement with the subject. A similar number of forms will be produced when the tense is changed into the past and the passive (with transitive verbs). Agreement shows itself to be a powerful element in Arabic grammar, and it also accounts for the complexity of the Arabic morphological system.

3.2.4. Demonstrative-Noun Agreement

English demonstrative pronouns agree with their head nouns in number.

this boy these boys	this girl these girls
that boy those boys	that girl those girls

French has a larger set of demonstratives to fulfill the requirement for agreement in number and gender.

ce garçon this boy	cette fille this girl
ces garcons these boys	ces filles these girls

Arabic demonstrative pronouns show greater requirements for agreement as they agree with their head noun in number (including the dual), gender, and case.

hāḏā-l-waladu
this (m, sg, NOM[1]) boy (NOM)

[1] Case has an overt word ending only with the dual demonstrative pronouns *hāḏāni* and *hātāni*; elsewhere, it is covert, i.e. not shown by an explicit word ending.

hāḏihī-l-bintu
this (f, sg, NOM) girl (NOM)

hāḏāni-l-waladān
these (m, dl, NOM) boys (NOM)

hāḏaini-l-waladaini
these (m, dl, ACC/GEN) boys (ACC/GEN)

hātāni-l-bintān
these (f, dl, NOM) girls (NOM)

hātaini-l-bintaini
these (f, dl, ACC/GEN) girls (ACC/GEN)

hāʾulāʾi-l-ʾawlādu
these (m/f, pl, NOM) boys (NOM)

hāʾulāʾ i-l-banātu
these (m/f, pl, NOM) girls (NOM)

3.2.5. Relative Pronoun-Antecedent Agreement

English relative pronouns agree with their antecedents in gender but not in number. The relative pronoun is marked only when the antecedent is non-human. Elsewhere, it has the same form.

the boy who plays
the girl who plays
the cat which plays

French relative pronouns do not have any agreement requirements, whether in number or gender. There is only the choice between *qui* and *que* according to the case of the word in the sentence (whether it is subject, object, or object of preposition).

la fille qui rit
the girl who laughs

la fille que j'aime
the girl whom I love

la fille à que je parle
the girl to whom I talk

Arabic relative pronouns, in contrast, agree with their controllers in number, gender and case.

ʾal-waladu-l-laḏī yalʿabu

the boy (NOM) who (m, sg, NOM[1]) plays

ʾal-bintu-l-latī talʿabu
the girl (NOM) who (m, sg, NOM) plays

ʾal-waladāni-l-laḏāni yalʿabāni
the boys (dl, NOM) who (m, dl, NOM) play

ʾal-waladaini-l-laḏaini yalʿabāni
the boys (dl, ACC/GEN) who (m, dl, ACC/GEN) play

ʾal-bintāni-l-latāni talʿabāni
the girls (dl, NOM) who (f, dl, NOM) play

ʾal-bintaini-l-lataini talʿabāni
the girls (dl, ACC/GEN) who (f, dl, ACC/GEN) play

ʾal-ʾawlādu-l-laḏīna yalʿabūna
the boys (pl, NOM) who (m, pl, NOM) play

ʾal-banātu-l-lātī (or ʾal-lāʾī) yalʿabna
the girls (pl, NOM) who (f, pl, NOM) play

3.2.6. Anaphora-Antecedent Agreement

Anaphoric pronouns include reflexives, reciprocals, "possessive anaphor[s]"[2] and pronominals.

English reflexives and possessives agree with their antecedent controllers in gender, number, and person.

I eat **my** food.
You eat **your** food.
He eats **his** food.
She eats **her** food.
It eats **its** food.
We eat **our** food.
They eat **their** food.

Arabic reflexives and possessives also agree with their antecedents in gender, number (singular, dual, or plural) and person.

ʾanā ʾākulu taʿāmī.
I eat my food.

[1] Case has an overt word ending only with the dual relative pronouns *al-laḏāni* and *al-latāni*; elsewhere, it is covert, i.e. not shown by an explicit word ending.
[2] Ellen Woolford, "More on the Anaphor Agreement Effect", *Linguistic Inquiry*, 30.2, 1999, 257-287, p. 274.

naḥnu naʾkulu taʿāmanā.
We eat our food.

ʾanta taʾkulu taʿāmaka.
You (m, sg) eat your (m, sg) food.

ʾanti taʾkulīna taʿāmaki.
You (f, sg) eat your (f, sg) food.

ʾantumā taʾkulāni taʿāmakumā.
You (m/f, dl) eat your (m/f, dl) food.

ʾantum taʾkulūna taʿāmakum.
You (m, pl) eat your (m, pl) food.

ʾantunna taʾkulna taʿāmakunna.
You (f, pl) eat your (f, pl) food.

huwa yaʾkulu taʿāmahu.
He eats his food.

hiya taʾkulu taʿāmahā.
She eats her food.

humā yaʾkulāni taʿāmahumā.
They (m, dl) eat their (m/f, dl) food.

humā taʾkulāni taʿāmahumā.
They (f, dl) eat their (m/f, dl) food.

hum yaʾkulūna taʿāmahum.
They (m, pl) eat their (m, pl) food.

hunna yaʾkulna taʿāmahunna.
They (f, pl) eat their (f, pl) food.

3.2.7. Predicate-Subǧect Agreement

In French, the predicative adjectives and nominals (adjectives and NPs that come after the copula) agree with the subject in number and gender. The following French sentences show agreement between the predicate (the target) and subject (the controller) in number and gender.

Il est petit. He is young.	Ils sont petits. They (m) are young.
Elle est petite. She is young.	Elles sont petites. They (f) are young.
Il est un étudiant. He is a student.	Ils sont des étudiants. They (m) are students.
Elle est une étudiante.	Elles sont des étudiantes.

She is a student. They (f) are students.

The following Arabic sentences also show agreement between the predicate and subject in number and gender.

huwa karīmun. He is generous.	hiya karīmatun. She is generous.
humā karīmāni. They (m, dl) are generous.	humā karīmatāni. They (f, dl) are generous.
hum kuramāʾu. They (m, pl) are generous.	hunna karīmātun. They (f, pl) are generous.
huwa ṭālibun. He is a student.	hiya ṭālibatun. She is a student.
humā ṭālibāni. They (m, dl) are students.	humā ṭālibatāni. They (f, dl) are students.
hum ṭullābun. They (m, pl) are students.	hunna ṭalibātun. They (f, pl) are students.

English adjectives are not marked for number or gender and so the predicative adjective does not agree with its subject. However, "a predicate nominal … must agree in number with the subject of its clause"[1].

They are scientists.
* They are a scientist.

3.2.8. Coordinated Word-Antecedent Agreement

In Arabic all words joined by a coordinating conjunction must agree only in case. This appears when coordinated words are nouns.

ʾal-bintu wa-l-ʾawlādu yuḥibbūna-t-tuffāḥ.
The girl (NOM) and the boys (NOM) like apples.

With pronouns, agreement does not appear as a case marking, because pronouns have fixed forms and are not overtly marked for case.

huwa wa hiya yuḥibbāni-t-tuffāḥ.
He and she like apples.

[1] Usama Mohamed Soltan, "A Contrastive and Comparative Syntactic Analysis of Deletion Phenomena in English and Standard Arabic", unpublished dissertation, Ain Shams University, 1996, p. 108.

English, in contrast, shows agreement only when the coordinated word is a pronoun.

They and **I** are teachers.
They invited **them** and **me**.

3.2.9. Targets with Double Controllers

French possessive anaphors have double controllers. They agree with their antecedent in person only, and they agree with the following head noun in gender and number.

Il marche avec son frère.
He walks with his brother.

Il marche avec ses frères.
He walks with his brothers.

Il marche avec sa sœur.
He walks with his sister.

Il marche avec ses sœurs.
He walks with his sisters.

Arabic includes a class of adjectives which has double controllers. They are called causative adjectives or *ʾan-na ṯu-s-sababiyy*, which describe something or somebody related to the head noun. These adjectives are always singular. They agree with the head noun in case and definiteness, and with the following noun in gender only[1].

hāḏā waladun karīmun ʾaḫūhu.
This boy (NOM) generous (m, sg, NOM) his brother.
This is a boy whose brother is generous.

hāḏā waladun karīmatun ʾuḫtuhu.
This boy (NOM) generous (f, sg, NOM) his sister.
This is a boy whose sister is generous.

raʾaitu-l-ʾawlāda-l-karīma ʾiḫwatuhum.
I saw the boys (ACC) the generous (m, sg, ACC) their brothers.
I saw the boys whose brothers are generous.

raʾaitu-l-ʾawlāda-l-karīmata ʾaḫawātuhum.
I saw the boys (ACC) the generous (f, sg, ACC) their sisters.
I saw the boys whose sisters are generous.

[1] Yūsuf Al-Ḥammādī, et al., *Al-Qawāʿidu-l-ʾAsāsiyyatu fī-n-Naḥwi wa-ṣ- ṣarf* [Basic Rules of Syntax and Morphology], Cairo: General Authority of National Publishing Houses, 1992, p. 136.

English, by contrast, does not show this type of agreement.

3.2.10. Anti-agreement

By anti-agreement I mean that the target takes the opposite features of the controller. This type is found in Arabic cardinal numbers and has two forms: anti-agreement in gender and anti-agreement in number.

Arabic cardinal numbers (from 3 to 10) which precede the head nouns show anti-agreement in gender. The rule for gender is that "masculine numbers are used with nouns whose singular is feminine, and feminine numbers are used with nouns whose singular is masculine."[1] The number takes a gender opposite to that of the base form of the noun, as the gender of non-human masculine nouns changes to the feminine in the plural form, such as:

kitāb	kutub
book (m, sg)	books (f, pl)

The numbers from 3 through 9 have the opposite gender of the noun when occurring alone as well as in the units (such as 3, 5, 7, 13, 35, 67). The following table shows examples:

No.	Masculine Number	Feminine Number
3	ṯalāṯu banāt three girls	ṯalāṯatu ʾawlād three boys
4	ʾarbaʿu banāt four girls	ʾarbaʾatu ʾawlād four boys
5	ẖamsu banāt five girls	ẖamsatu ʾawlād five boys

[1] Raja T. Nasr, *The Structure of Arabic: From Sound to Sentence*, Beirut: Librairie du Liban, 1967, p. 121.

6	sittu banāt six girls	sittatu ʾawlād six boys
7	sabʿu banāt seven girls	sabʿatu ʾawlād seven boys
8	ṯamanī banāt eight girls	ṯamaniyatu ʾawlād eight boys
9	tisʿu banāt nine girls	tisʿatu ʾawlād nine boys
10	ʿašru banāt ten girls	ʿašratu ʾawlād ten boys

Table 11. Gender anti-agreement

The Arabic cardinal numbers from 11 and above show anti-agreement in number. While the cardinal number is evidently plural, the head noun is always singular, in contrast with the numbers 3 through 9 where the head noun is always plural.

In the cardinal numbers 11 and 12, both elements of the number, i.e. elements in the units and the tens, agree with the head noun in gender.

No.	**1st Part m. / 2nd part m.**	**1st Part f. / 2nd part f.**
11	ʾaḥada ʿašara waladan eleven boy	ʾiḥdā ʿašrata bintan eleven girl
12	ʾiṯnā ʿašara waladan twelve boy	ʾiṯnatā ʿašrata bintan twelve girl

Table 12. Number anti-agreement

In the numbers 13 through 19, the first element has the opposite gender while the second element positively agrees in gender with the head noun. The head noun is again invariably singular.

No.	1st Part m. / 2nd part f.	1st Part f. / 2nd part m.
13	ṯalāṯa ʿašrata bintan thirteen girl	ṯalāṯata ʿašara waladan thirteen boy
14	ʾarbaʿa ʿašrata bintan forteen girl	ʾarbaʿata ʿašara waladan forteen boy
15	ẖamsa ʿašrata bintan fifteen girl	ẖamsata ʿašara waladan fifteen boy
16	sitta ʿašrata bintan sixteen girl	sittata ʿašara waladan sixteen boy
17	sabʿa ʿašrata bintan seventeen girl	sabʿata ʿašara waladan seventeen boy
18	ṯamāniya ʿašrata bintan eighteen girl	ṯamaniyata ʿašara waladan eighteen boy
19	tisʿa ʿašrata bintan nineteen girl	tisʿata ʿašara waladan nineteen boy

Table 13. Gender and number anti-agreement

The numbers from 20 and above have no gender-related variations. However, whenever the numbers 1 and 2 are in the units, they agree in gender with the head noun, and whenever the numbers 3 through 9 are in the units they take the opposite gender. The head noun is always singular.

Chapter Four
Agreement and Cross-Language Redundancies

Agreement is a cohesive element in the sentence structure which relates constituents to each other. However, the agreement system differs greatly from one language to the other, a matter that requires a lot of attention during translation. In this section I will explain the agreement system in standard Arabic (SA) and in English in detail.

4.1. Agreement in Arabic

Agreement is a very powerful feature in Arabic. Standard Arabic is "an agreement language, with a rich and complex agreement system"[1] that interacts with different syntactic elements. The diversity of inflections in Arabic morphology is due in great parts to the requirements of agreement. The Arabic language has thirteen agreement features in five grammatical categories distributed as follows:

I. Agreement in gender (masculine, and feminine)

II. Agreement in number (singular, dual, and plural)

III. Agreement in definiteness (definite, and indefinite)

IV. Agreement in case (nominative, accusative, and genitive)

V. Agreement in person (first person, second person, and third person)

4.1.1. Agreement Categories

4.1.1.1. Gender

Gender is the system by which a language expresses the natural or grammatical sex of objects. Natural sex means the natural classification of animate objects into males and females such as *man/woman* and *horse/mare*. Grammatical sex means the lexical classification of words into feminine and masculine regardless of their natural sex, such as the Arabic words *kitāb* (book, masculine) /*kurrāsah* (notebook, feminine). The importance of gender stems from the fact that a "noun belonging to a

[1] Abdelkader Fassi Fehri, *Issues in the Structure of Arabic Clauses and Words*, Dordrecht: Kluwer Academic Publishers, 1993, p. xi.

particular gender determines the choice among a set of alternative "agreeing" forms … of morphemes or words"[1]. The way gender triggers agreement is interesting. We notice that languages which employ grammatical gender, like Arabic and French, make gender-related agreement more than languages which have no grammatical gender, like English.

Arabic gender has two features: feminine and masculine. Basically, "gender in Arabic is grammatical, not natural".[2] Therefore all Arabic nouns (whether singular, dual or plural; human or non-human; animate or inanimate) are classified according to gender into feminine and masculine. "Every noun in SA is either masculine or feminine"[3].

Nouns, with regard to gender marking, can be broadly divided into animate (whose gender is natural) and inanimate (whose gender is grammatical). The default unmarked form of the animate nouns is the masculine. In most cases the feminine can be derived from a masculine stem by attaching a feminine morpheme to the end of the word:

qiṭṭ cat (m)	qiṭṭah cat (f)
ṭālib student (m)	ṭālibah student (f)
mudarris teacher (m)	mudarrisah teacher (f)

Sometimes the feminine is not derived from the masculine but has a completely different form:

walad boy (m)	bint girl (f)
raǧul man (m)	ʾimraʾah woman (f)
dīk cock (m)	daǧāǧah chicken (f)

[1] Joseph H. Greenberg, "Gender Markers", in Joseph H. Greenberg, ed., *Universals of Human Language*, Stanford, California: Stanford University Press, 1978, p. 49.
[2] Raja T. Nasr, *The Structure of Arabic: From Sound to Sentence*, Beirut: Librairie du Liban, 1967, p. 100.
[3] Reima Al-Jarf, "Grammatical agreement errors in L1/L2 translations. IRAL: *International Review of Applied Linguistics in Language Teaching*, 38.1, 2000, 1-15.

The animate nouns can be further classified into human and non-human nouns. Human nouns preserve their gender in singular, dual, and plural forms.

ṭālib student (m, sg)	ṭālibān students (m, dl)	ṭullāb students (m, pl)
ṭālibah student (f, sg)	ṭālibatān students (f, dl)	ṭālibāt students (f, pl)

Non-human nouns, in contrast, preserve their gender in singular and dual forms only. In the plural form, they invariably take the feminine gender.

qiṭ cat (m)	qiṭṭān cats (m, dl)	qiṭaṭ cats (f, pl)
dīk cock (m)	dīkān cocks (m, dl)	diyūk cocks (f, pl)

The inanimate nouns can be either masculine or feminine. The feminine is not derived from the masculine nor does it have a masculine counterpart. The gender of inanimate nouns is considered a lexical property of the word. For each inanimate noun, the dictionary specifies whether it is masculine or feminine.

qamar moon (m)	šams sun (f)
kitāb book (m)	kurrāsah notebook (f)
ḥağar stone (m)	saǧrah rock (f)

Like animate, non-human nouns, the grammatical gender of inanimate nouns is preserved in singular and dual forms, but they are invariably feminine when they are changed into the plural form.

kitāb book (m)	kitābān books (m, dl)	kutub books (f, pl)

4.1.1.2. Number

Number is the system by which a language denotes the singularity, duality, or plurality of objects. Arabic number has three features: singular, dual, and plural. The singular means 'one', the dual means 'two' and the plural means "more than two"[1]. The default unmarked form of the noun is the singular; the dual and the plural are derived from the singular. The dual is formed by suffixing a morpheme (*-ān* or *-ain* depending on the case) to the stem noun whether it is designating human, non-human, animate, non-animate, feminine or masculine objects.

mudarris teacher (m)	mudarrisān teachers (m, dl)
mudarrisah teacher (f)	mudarrisatān teachers (f, dl)
kalb dog (m)	kalbān dogs (m, dl)
kalbah dog (f)	kalbatān dogs (f, dl)
kitāb book (m)	kitābān books (m, dl)
kurrāsah notebook (f)	kurrāsatān notebooks (f, dl)

The plural of human feminine nouns is invariably formed by suffixing the regular feminine plural[2] morpheme (*-āt*) to the word stem.

mudarrisah teacher (f)	mudarrisāt teachers (f, pl)
ṭālibah student (f)	ṭālibāt students (f, pl)

For non-human feminine nouns, the plural can be formed either by suffixing the regular feminine plural morpheme (*-āt*) to the word stem, or it can have a special irregular form[3].

baqarah cow (f)	baqarāt cows	baqar cows

[1] Abbas Hasan. *An-Naḥwu-l-Wāfī* [Comprehensive Grammar], 14th ed., Cairo, Egypt: Dar Al-Maʿarif, 1999, p. 137.
[2] Called by Arabic grammarians: ǧamʿu-l-muʾannathu-s-sālim (or sound feminine plural).
[3] Called by Arabic grammarians: ǧamʿu taksīr (or broken plural).

The plural of human masculine nouns is sometimes formed by suffixing the regular masculine plural[1] morpheme (*-ūn* or *-īn* depending on the word case).

mudarris teacher	mudarrisūn teachers
muhandis engineer	muhandisūn engineers

In other cases, it has an irregular plural form.

ṭālib student	ṭullāb students
ṭabīb physician	aṭibbaʾ physicians

For non-human masculine nouns, the plural is formed by changing the word form.

dīk cock	diyūk cocks
kitāb book	kutub books

It must be noted that some exceptions may exist along the borderlines of the rules mentioned above. Moreover, Arabic includes a class of nouns which has no singular variant.

nās
people

nisāʾ
women

[1] Called by Arabic grammarians: ǧamʿu-l-muḏakkari-s-sālim (or sound masculine plural).

4.1.1.3. Definiteness

Definiteness is the system by which a language expresses the speaker's and hearer's familiarity or unfamiliarity with the object spoken about. Definiteness has two features: definite and indefinite. Arabic indefinite nouns are introduced with the bare form without using any article. The definite article in Arabic is *ʾal-* which is similar to the English article *the*. The Arabic definite article may be realized as it is or "assimilated in the word it is connected with,"[1] depending on the initial consonant of the word. When it is assimilated the sound /l/ is realized as a gemination of the first consonant of the word the article is attached to. Assimilation occurs when the noun begins with any of the following letters: t, ṯ, d, ḏ, r, z, s, š, ṣ, ḍ, ṭ, ẓ, l and n. In all other cases it is pronounced as it is and not assimilated.

Examples of clear - *ʾal*:

ʾal-qamar	ʾal-kitāb
the moon	the book
ʾal-bāb	ʾal-māʾ
the door	the water

Examples of assimilated - *ʾal*:

ʾaš-šams	ʾaṭ-ṭarīq
the sun	the road
ʾas-samāʾ	ʾan-nās
the sky	the people

4.1.1.4. Case

Case is usually realized as a word ending that indicates the grammatical functional relationship of a word or phrase to other constituents of the sentence. Arabic case has three main[2] features: nominative, accusative, and genitive. Case may be realized as an overt word ending, or it may be covert with no apparent effect on the word.

[1] Raja T. Nasr, *The Structure of Arabic: From Sound to Sentence*, Beirut: Librairie du Liban, 1967, p. 181.

[2] There is a fourth case in Arabic called *jazm* (jussive), which affects only verbs in certain contexts.

The nominative (NOM) case applies, most notably, to the subject, and can be realized as *-u* for the singular, *-ān* for the dual, and *-ūn* for the plural.

ǧāʾa-l-mudarrisu.
came the teacher (sg, NOM)
The teacher came.

ǧāʾa-l-mudarrisāni.
came the teachers (dl, NOM)
The teachers came.

ǧāʾa-l-mudarrisūna.
came the teachers (pl, NOM)
The teachers came.

The accusative (ACC) case applies, among other things, to the object, and can be realized as *-a* for the singular, *-aṁ* for the dual, and *-īn* for the plural.

raʾaitu-l-mudarrisa.
saw I the teacher (sg, ACC)
I saw the teacher.

raʾaitu-l-mudarrisaini.
saw I the teachers (dl, ACC)
I saw the teachers.

raʾaitu-l-mudarrisīna.
saw I the teachers (pl, ACC)
I saw the teachers.

The genitive (GEN) case applies, among other things, to the object of preposition, and can be realized as *-i* for the singular, *-aṁ* for the dual, and *-īn* for the plural.

ḏahabtu ʾilā-l-mudarrisi.
went I to the teacher (sg, GEN)
I went to the teacher.

ḏahabtu ʾilā-l-mudarrisaini.
went I to the teachers (dl, GEN)
I went to the teachers.

ḏahabtu ʾilā-l-mudarrisīna.
went I to the teachers (pl, GEN)
I went to the teachers.

In the accusative and genitive cases there are exceptions in classes of nouns such as the regular feminine plural, diptote nouns[1], and the five nouns[2].

4.1.1.5. Person

Person is the system by which a language expresses the different speech acts of speaker, hearer, and other. The pronoun system is not universally unified, but varies from one language to another, some using more or less personal pronouns than others. This variation can be explained in terms of the "roles in the speech act that each language considers to be of sufficient importance to mark by a separate lexical form"[3].

Personal pronouns universally have three features: first person, second person and third person. Arabic person is expressed by the list of pronouns in the language. Arabic uses 12 different lexical items to cover 18 personal pronouns.

Person	**Gender**	**Singular**	**Dual**	**Plural**
First	Masculine	ʾanā	naḥnu	
	Feminine			
Second	Masculine	ʾanta	ʾantumā	ʾantum
	Feminine	ʾanti		ʾantunna
Third	Masculine	huwa	humā	hum
	Feminine	hiya		hunna

Table 14. Arabic pronoun system

In his interesting study of personal pronouns, Ingram[4] classified the personal pronoun systems of the world languages according to their person roles, ignoring the features of gender and case. On Ingram's scale, Arabic is represented[5] as an eight-person system, with the following structure:

[1] In Arabic *al-mamnūʾ mina-ṣ-ṣarf* (nouns barred from complete inflection).
[2] In Arabic *al-ʾasmāʾ u-l-ḫamsah* (they are: *abū, aḫū, ḥamū, fū*, and *ḏū*).
[3] David Ingram, "Personal Pronouns", in Joseph H. Greenberg, ed., *Universals of Human Language*, Stanford, California: Stanford University Press, 1978, p. 215.
[4] Ibid., p. 213.
[5] Ibid., p. 244.

I		we
thou	you-2	you
he	they-2	they

Table 15. Arabic pronoun system according to Ingram

This means that Arabic has a dual number for the pronouns *thou* (i.e. *ʾanta*, the dual Arabic pronoun is *ʾantumā*) and *he* (i.e. *huwa*, the dual Arabic pronoun is *humā*) but no dual for the pronoun *I*.

4.1.2. Agreement Degrees

In the previous chapter I discussed ten types of agreement. Standard Arabic has all types except the first one, i.e. article-noun agreement. Some types require more agreement features than the others. So, in this chapter I will make a classification according to the degree of agreement required.

4.1.2.1. Agreement in Gender, Number, Definiteness, and Case

This is the largest number of agreement features that can be combined together. It is realized only in adjectives when they modify a noun of a human referent. In this case adjectives must agree with nouns in gender, number, definiteness, and case.

ʾal-mudīru-n-nāğiḥu
the manager (m, sg, NOM) the successful (m, sg, NOM)
the successful manager

ʾal-mudīratu-n-nāğiḥatu
the manager (f, sg, NOM) the successful (f, sg, NOM)
the successful manager

ʾal-mudīrāni-n-nāğiḥāni
the managers (m, dl, NOM) the successful (m, dl, NOM)
the successful managers

ʾal-mudīratāni-n-nāğiḥatāni
the managers (f, dl, NOM) the successful (f, dl, NOM)
the successful managers

ʾal-mudīrūna-n-nāğiḥūna
the managers (m, pl, NOM) the successful (m, pl, NOM)
the successful managers

ʾal-mudīrātu-n-nāğiḥātu
the managers (f, pl, NOM) the successful (f, pl, NOM)
the successful managers

As we notice here, adjectives agree with the nouns they modify in gender (masculine [m] or feminine [f]), number (singular [sg], dual [dl], or plural [pl]), definiteness (definite or indefinite) and case (nominative [NOM], accusative [ACC], or genitive [GEN]). This means that if we expanded our example to make account for other possible cases, we shall have 14 variations as listed below. Moreover, this number can be doubled when considering removing the definite article from each.

1. ʾan-nāğiḥu (m,sg,NOM)
2. ʾan-nāğiḥa (m,sg,ACC)
3. ʾan-nāğiḥi (m,sg,GEN)

4. ʾan-nāğiḥatu (f,sg,NOM)
5. ʾan-nāğiḥata (f,sg,ACC)
6. ʾan-nāğiḥati (f,sg,GEN)

7. ʾan-nāğiḥāni (m,dl,NOM)
8. ʾan-nāğiḥaini (m,dl,ACC/GEN)
9. ʾan-nāğiḥatāni (f,dl,NOM)
10. ʾan-nāğiḥataini (f,dl,ACC/GEN)

11. ʾan-nāğiḥūna (m,pl,NOM)
12. ʾan-nāğiḥīna (m,pl,ACC/GEN)

13. ʾan-nāğiḥātu (f,pl,NOM)
14. ʾan-nağiḥāti (f,dl,ACC/GEN)

However, if the head noun is a "broken plural of a non-human referent, the target is feminine" [1] and singular. This means that the target, i.e. the adjective, will not agree in number with the controller, i.e. the head noun.

ʾal-kutubu-l-mufīdatu
the books (f, pl, NOM) the useful (f, sg, NOM)
the useful books

4.1.2.2. Agreement in Gender, Number, and Person

This degree of agreement is realized in two cases: in verb-subject agreement and in the agreement of anaphoric pronoun (whether reflexive or possessive) with the antecedent.

a. Verb-Subject Agreement

[1] Reima Al-Jarf, "Grammatical agreement errors in L1/L2 translations. IRAL: *International Review of Applied Linguistics in Language Teaching*, 38.1, 2000, 1-15.

When the target is a verb and the controller is a subject and the verb follows the subject[1], then the verb must agree with the subject in gender, number, and person (first, second, or third).

ʾanā ašrabu
I drink (m/f)

naḥnu našrabu
we drink (m/f)

ʾanta tašrabu you (m, sg) drink (m, sg)	ʾanti tašrabīna you (f, sg) drink (f, sg)
ʾantumā tašrabāni you (m/f, dl) drink (m/f, dl)	
ʾantum tašrabūna you (m, pl) drink (m, pl)	ʾantunna tašrabna you (f, pl) drink (f, pl)
huwa yašrabu he drinks	hiya tašrabu she drinks
humā yašrabāni they (m, dl) drink (m, dl)	humā tašrabāni they (f, dl) drink (f, dl)
hum yašrabūna they (m, pl) drink (m, pl)	hunna yašrabna they (f, pl) drink (f, pl)

Even when the verb tense is changed into the past or the voice changed into the passive or the mood changed into the jussive (imperative), the same agreement degree is required.

If the subject consists of two or more conjoined NPs, the verb must still agree with them in number and gender. If the conjoined NPs have different genders, the verb takes the masculine gender, which is also considered the default unmarked gender.

fāṭimah wa salmā ḏahabatā ʾilā-l-madrasati
Faṭimah and Salmā went (f, dl) to school.

aḥmad wa maḥmūd ḏahabā ʾilā-l-madrasati
Aḥmad and Maḥmūd went (m, dl) to school.

ʾal-waladu wa-l-bintu ḏahabā ʾilā-l-madrasati
The boy and the girl went (m, dl) to school.

ʾal-ʾawlādu wa-l-banātu ḏahabū ʾilā-l-madrasati
The boys and the girls went (m, pl) to school.

[1] This structure is known in Arabic grammar as the nominal sentence, which consists of a subject and a predicate.

A complete picture of the morphological varieties of the Arabic verb can be drawn by the following table. It shows how verbs in Arabic are inflected to accommodate gender, number, and person agreement.

Person	Preterit		Present		Passive (Past)		Passive (Present)		Imperative	
ʾanā [I]	ḍarabtu	ضَرَبْتُ	ʾaḍribu	أَضْرِبُ	ḍuribtu	ضُرِبْتُ	ʾuḍrabu	أُضْرَبُ		
naḥnu [we]	ḍarabnā	ضَرَبْنا	naḍribu	نَضْرِبُ	ḍuribnā	ضُرِبْنا	nuḍrabu	نُضْرَبُ		
ʾanta [you (m,sg)]	ḍarabta	ضَرَبْتَ	taḍribu	تَضْرِبُ	ḍuribta	ضُرِبْتَ	tuḍrabu	تُضْرَبُ	ʾiḍrib	اضرب
ʾanti [you (f,sg)]	ḍarabti	ضَرَبْتِ	taḍribīna	تَضْرِبِين	ḍuribti	ضُرِبْتِ	tuḍrabina	تُضْرَبِين	ʾiḍribī	اضربي
ʾantumā [you (m/f,dl)]	ḍarabtumā	ضَرَبْتُما	taḍribāni	تَضْرِبان	ḍuribtumā	ضُرِبْتُما	tuḍrabāni	تُضْرَبان	ʾiḍribā	اضربا
ʾantum [you (m,pl)]	ḍarabtum	ضَرَبْتُم	taḍribūna	تَضْرِبُون	ḍuribtum	ضُرِبْتُم	tuḍrabūna	تُضْرَبُون	ʾiḍribū	اضربوا
ʾantunna [you (f,pl)]	ḍarabtunna	ضَرَبْتُنَّ	taḍribna	تَضْرِبن	ḍuribtunna	ضُرِبْتُن	tuḍrabna	تُضْرَبن	ʾiḍribna	اضربن
huwa [he]	ḍaraba	ضَرَبَ	yaḍribu	يَضْرِبُ	ḍuriba	ضُرِبَ	yuḍrabu	يُضْرَبُ		
hiya [she]	ḍarabat	ضَرَبَتْ	taḍribu	تَضْرِبُ	ḍuribat	ضُرِبَتْ	tuḍrabu	تُضْرَبُ		
humā (m.) [they (dl)]	ḍarabā	ضَرَبا	yaḍribāni	يَضْرِبان	ḍuribā	ضُرِبا	yuḍrabāni	يُضْرَبان		
humā (f.) [they (dl)]	ḍarabatā	ضَرَبَتا	taḍribāni	تَضْرِبان	ḍuribatā	ضُرِبتا	tuḍrabāni	تُضْرَبان		
hum [they (m,pl)]	ḍarabū	ضَرَبُوا	yaḍribūna	يَضْرِبُون	ḍuribū	ضُرِبوا	yuḍrabūna	يُضْرَبُون		
hunna [they (f,pl)]	ḍarabna	ضَرَبْنَ	yaḍribna	يَضْرِبن	ḍuribna	ضُرِبْن	yuḍrabna	يُضْرَبن		

Table 16. Possible inflections of the verb *ḍaraba* (to hit) [1]

[1] Adapted from Raja T. Nasr, *The Structure of Arabic: From Sound to Sentence*, Beirut: Librairie du Liban, 1967, p. 77.

b. Reflexive Pronoun-Antecedent Agreement

Arabic reflexive pronouns agree with their antecedents in gender, number, and person.

ʾanā ʾuḥibbu nafsī.
I love myself.

naḥnu nuḥibbu anfusanā.
We love ourselves.

ʾanta tuḥibbu nafsaka.
You (m, sg) love yourself (m, sg).

ʾanti tuḥibbīna nafsaki.
You (f, sg) love yourself (f, sg).

ʾantumā tuḥibbāni nafsaikumā.
You (m/f, dl) love yourselves (m/f, dl).

ʾantum tuḥibbūna anfusakum.
You (m, pl) love yourselves (m, pl).

ʾantunna tuḥbibna anfusakunna.
You (f, pl) love yourselves (f, pl).

huwa yuḥibbu nafsahu.
He loves himself.

hiya tuḥibbu nafsahā.
She loves herself.

humā yuḥibbāni nafsaihimā.
They (m, dl) love themselves (m/f, dl).

humā tuḥibbāni nafsaihimā.
They (f, dl) read their (m/f, dl) book.

hum yuḥibbūna anfusahum.
They (m, pl) love themselves (m, pl).

hunna yuḥbibna anfusahunna.
They (f, pl) love themselves (f, pl).

c. Possessive Pronoun-Antecedent Agreement

Arabic possessive pronominals agree with their antecedents in gender, number, and person.

ʾanā ʾaqraʾu kitābī.
I read my book.

naḥnu naqraʾu kitābanā.
We read our book.

ʾanta taqraʾu kitābaka.
You (m, sg) read your (m, sg) book.

ʾanti taqraʾīna kitābaki.
You (f, sg) read your (f, sg) book.

ʾantumā taqraʾāni kitābakumā.
You (m/f, dl) read your (m/f, dl) book.

ʾantum taqraʾūna kitābakum.
You (m, pl) read your (m, pl) book.

ʾantunna taqraʾna kitābakunna.
You (f, pl) read your (f, pl) book.

huwa yaqraʾu kitābahu.
He reads his book.

hiya taqraʾu kitābahā.
She reads her book.

humā yaqraʾāni kitābahumā.
They (m, dl) read their (m, dl) book.

humā taqraʾāni kitābahumā.
They (f, dl) read their (f, dl) book.

hum yaqraʾūna kitābahum.
They (m, pl) read their (m, pl) book.

hunna yaqraʾna kitābahunna.
They (f, pl) read their (f, pl) book.

4.1.2.3. Agreement in Gender, Number, and Case

This is the most common degree of agreement. It is found in demonstrative pronoun-noun and relative pronoun-antecedent agreement.

a. Demonstrative Pronoun-Noun Agreement

Demonstratives agree with the head nouns in gender (masculine or feminine), number (singular, dual, or plural) and case (nominative, accusative, or genitive), whether they refer to near or far objects. It must be noted that with demonstrative pronouns, case is always covert except in the dual.

Near

hāḏa-l-waladu
this (m, sg, NOM) boy (m, sg, NOM)

hāḏihi-l-bintu
this (f, sg, NOM) girl (f, sg, NOM)

hāḏāni[1]-l-waladāni
these (m, dl, NOM) boys (m, dl, NOM)

hātāni[2]-l-bintāni
these (f, dl, NOM) girls (f, dl, NOM)

hāʾulāʾi-l-ʾawlādu
these (m/f, pl, NOM) boys (m, pl, NOM)

hāʾulāʾi-l-banātu
these (m/f, pl, NOM) girls (f, pl, NOM)

Far

ḏālika-l-waladu
that (m, sg, NOM) boy (m, sg, NOM)

tilka-l-bintu
this (f, sg, NOM) girl (f, sg, NOM)

ḏānika[3]-l-waladāni
those (m, dl, NOM) boys (m, dl, NOM)

tānika[4]-l-bintāni
those (f, dl, NOM) girls (f, dl, NOM)

ʾulāʾika-l-ʾawlādu
those (m/f, pl, NOM) boys (m, pl, NOM)

ʾulāʾika-l-banātu
those (m/f, pl, NOM) girls (f, pl, NOM)

b. Relative Pronoun-Antecedent Agreement

Relative Pronouns agree with their antecedents in gender, number, and case. Case is covert in all relative pronouns except in the dual.

ʾal-waladu-l-laḏī yalʿabu
the boy (m, sg, NOM) who (m, sg, NOM) plays

ʾal-bintu-l-latī talʿabu
the girl (f, sg, NOM) who (f, sg, NOM) plays

ʾal-waladāni-l-laḏāni5 yalʿabāni
the boys (m, dl, NOM) who (m, dl, NOM) play

[1] In the accusative and genitive *haḏāni* becomes *haḏaini*.
[2] In the accusative and genitive *hatāni* becomes *hataini*.
[3] In the accusative and genitive *ḏānika* becomes *ḏainika*.
[4] In the accusative and genitive *tānika* becomes *tainika*.
[5] In the accusative and genitive *al-laḏāni* becomes *al-laḏaini*.

ʾal-bintāni-l-latāni[1] talʿabāni
the girls (f, dl, NOM) who (f, dl, NOM) play

ʾal-ʾawlādu-l-laḏīna yalʿabūna
the boys (m, pl, NOM) who (m, pl, NOM) play

ʾal-banātu-l-lātī[2] talʿabna
the girls (f, pl, NOM) who (f, pl, NOM) play

The non-human plural head nouns have singular feminine demonstrative and relative pronouns.

hāḏihi-l-diyuk
This (f, sg) cocks

ʾal-diyuku-l-latī
The cocks which (f, sg)

The Arabic relative pronoun system also includes two pronouns which are not marked either for gender or number. They are *man* (similar to English *who*) for humans and *ma* (similar to English *which*) for non-humans.

4.1.2.4. Agreement in Gender and Number

When a noun or an adjective is used as a predicate it agrees with its subject in gender and number.

Predicative Nominals

huwa ṭabībun.
He (m, sg) is a physician (m, sg).

hiya ṭabībatun.
She (f, sg) is a physician (f, sg).

humā ṭabībāni.
They (m, dl) are physicians (m, dl).

[1] In the accusative and genitive *al-latāni* becomes *al-lataini*.
[2] Or *al-banātu-l-lāʾī*.

humā ṭabībatāni.
They (f, dl) are physicians (f, dl).

hum aṭibbāʾun.
They (m, pl) are physicians (m, pl).

hunna ṭabībātun.
They (f, pl) are physicians (f, pl).

Predicative Adjectives

huwa karīmun.
He (m, sg) is generous (m, sg).

hiya karīmatun.
She (f, sg) is generous (f, sg).

humā karīmāni.
They (m, dl) are generous (m, dl).

humā karīmatāni.
They (f, dl) are generous (f, dl).

hum kuramāʾu.
They (m, pl) are generous (m, pl).

hunna karīmātun.
They (f, pl) are generous (f, pl).

Arabic uses different rules of agreement in comparatives and superlatives[1]. In the comparative form, the adjective does not agree with the subject, but has the default form of masculine and singular.

ʾal-waladu afḍalu min aḫīhi.
The boy is better (m, sg) than his brother.

ʾal-bintu afḍalu min uḫtihā.
The girl is better (m, sg) than her sister.

ʾal-ʾawlādu afḍalu min iḫwatihim.
The boys are better (m, sg) than their brothers.

ʾal-banātu afḍalu min aḫawātihinna.
The girls are better (m, sg) than their sisters.

In the superlative form, when the adjective is used predicatively, it agrees with the subject in number and gender.

[1] The rules related to the comparative and the superlative are adapted from Yūsuf Al-Ḥammādī, et al., *Al-Qawāʿidu-l-ʾAsāsiyyatu fī-n-Naḥwi wa-ṣ- ṣarf* [Basic Rules of Syntax and Morphology], Cairo: General Authority of National Publishing Houses, 1992, pp. 216-217.

hāḏā-l-waladu huwa-l-ʾafḍalu.
This boy (m, sg) is the best (m, sg).

hāḏihī-l-bintu hiya-l-fuḍlā.
This girl (f, sg) is the best (f, sg).

hāḏāni-l-waladāni humā-l-ʾafḍalāni.
These boys (m, dl) are the best (m, dl).

hātāni-l-bintāni humā-l-fuḍliāni.
These girls (f, dl) are the best (f, dl).

hāʾulāʾi-l-ʾawlādu humu-l-ʾafḍalūna.
These boys (m, pl) are the best (m, pl).

hāʾulāʾi-l-banātu hunna-l-fuḍlayātu.
These girls (f, pl) are the best (f, pl).

When the superlative is used attributively, i.e. the superlative is followed by a definite noun (modified by the definite article - *ʾal*), then the two options are available: agreement in number and gender, and having the default form of masculine and singular all through.

hāḏā-l-waladu huwa ʾafḍalu-ṭ-ṭullābi.
This boy (m, sg) is the best (m, sg) student.

hāḏihī-l-bintu hiya fuḍlā/ʾafḍalu-ṭ-ṭalibāti.
This girl (f, sg) is the best (f, sg)/best (m, sg) student.

hāḏāni-l-waladāni humā ʾafḍalā/ʾafḍalu-ṭ-ṭullābi.
These boys (m, dl) are the best (m, dl)/best (m, sg) students.

hātāni-l-bintāni humā fuḍlayā/ʾafḍalu-ṭ-ṭalibāti.
These girls (f, dl) are the best (f, dl)/best (m, sg) students.

hāʾulāʾi-l-ʾawlādu hum ʾafḍalū/ʾafḍalu-ṭ-ṭullābi.
These boys (m, pl) are the best (m, pl)/best (m, sg) students.

hāʾulāʾi-l-banātu hunna fuḍlayātu/ʾafḍalu-ṭ-ṭalibāti.
These girls (f, pl) are the best (f, pl)/best (m, sg) students.

If the noun following the superlative, however, is not preceded by the definite article – *ʾal*, it must be masculine and singular in all cases.

4.1.2.5. Agreement in Gender Only

Arabic has a relatively free word order. It can have SVO (*ʾal-waladu ʾakala-t-tuffaḥata*: the boy ate the apple), VSO (*ʾakala-l-waladu-t-tuffaḥata*: ate the boy the apple), OVS (*ʾat-tuffaḥata ʾakalaha-l-waladu*: the apple ate the boy) and VOS (*ʾakala-t-tuffahata-l-waladu*: ate the apple the boy).

The preferred word order, however, is VSO, in which case the verb agrees with the subject in gender only.

šariba-l-waladu-l-labana.	šaribati-l-bintu-l-labana.
drank (m) the boy the milk.	hit (f) the girl the milk.
The boy drank the milk.	The girl drank the milk.
šariba-l-waladāni-l-labana.	šaribati-l-bintāni-l-labana.
drank (m) the boys (dl) the milk.	drank (f) the girls (dl) the milk.
The boys drank the milk.	The girls drank the milk.
šariba-l-ʾawlādu-l-labana.	šaribati-l-banātu-l-labana.
drank (m) the boys (pl) the milk.	drank (f) the girls (pl) the milk.
The boys drank the milk.	The girls drank the milk.

If the subject consists of two conjoined NPs, the verb agrees in gender with the first NP, i.e. the NP that immediately follows the verb.

ḏahaba-l-waladu wa-l-bintu ʾilā-l-madrasati.
went (m, sg) the boy and the girl to the school.
The boy and the girl went to school.

ḏahabati-l-bintu wa-l-waladu ʾilā-l-madrasati.
went (f, sg) the girl and the boy to the school.
The girl and the boy went to school.

4.1.2.6. Agreement in Case Only

Agreement in case only is realized in coordination. When a word or phrase is used after a coordinating conjunction (such as *wa* [and] or *ʾau* [or]) it must agree with its antecedent in case, whether nominative (NOM), accusative (ACC), or genitive (GEN).

ʾal-waladu wa-l-bintu yalʿabāni.
The boy (NOM) and the girl (NOM) play.

raʾaitu-l-walada wa-l-binta.
I saw the boy (ACC) and the girl (ACC).

ḏahabtu ʾilā-l-waladi wa-l-binti.
I went to the boy (GEN) and the girl (GEN).

4.2. Agreement in English

The agreement system in English is less diversified and less varied than the Arabic system. Agreement in English is very limited both in features and types.

4.2.1. Agreement Categories

4.2.1.1. Gender

Gender has three features: feminine, masculine, and neuter. There are nouns that can be classified as feminine such as *woman*, *girl*, and *aunt*; nouns that can be classified as masculine such as *man*, *boy*, and *uncle*; and nouns that can be classified as neuter such as *cat*, *dog*, and *stone*.

However, the major bulk of English nouns cannot be classified according to gender. Words such as *student*, *doctor*, *engineer*, *fighter*, *player*, *teacher*, etc. are common nouns and cannot be classified as either masculine, feminine, or neuter. These nouns are regarded by some linguists as having a "common"[1] gender because they refer to either sex. They have "dual"[2] gender specifications as they can be referred to with the pronoun *he* or *she*.

Grammatical gender requires that nouns, as well as adjectives, be inflected to indicate whether they are masculine, feminine, or neuter. This kind of inflection is not found in English. English does not make the broad classification of nouns according to gender, yet it has some lexical items which refer to male, female, or sexless nouns.

In English there is a clear and close "connection between the biological category 'sex' and the grammatical category 'gender'"[3]. This leads us to the assumption that nouns in the English language are not marked for gender, or as Palmer puts it: "English has no gender."[4] Therefore, the male-female pairs of words (such as *man/woman* and *boy/girl*) are marked for "sex, not gender"[5]. The features attached to such words are lexical features denoting sex rather than grammatical features

[1] J. C. Nesfield and F. T. Wood, *Manual of English Grammar and Composition*, London: Macmillian, 1964, p. 24.
[2] Randolph Quirk and Sidney Greenbaum, *A University Grammar of English*, Harlow, Essex: Longman, 1973, p. 90.
[3] Ibid., p. 89.
[4] Frank Palmer, *Grammar*, Middlesex: Penguin Books, 1971, p. 189.
[5] Ibid.

denoting gender. I will continue, however, to speak about gender in English, while, in fact, referring to this sex distinction.

A further distinction in English is between gender-specific nouns and common-gender nouns[1]. Gender-specific nouns denote either male or female entities such as *husband*/*wife*, *father*/*mother*, and *man/woman*. By contrast common-gender nouns apply to either sex such as *spouse*, *parent*, and *person*.

The suffix *-ess* is used in some nouns to make male-female distinction such as in *prince/princess*, *waiter/waitress*, *actor/actress*, *ambassador/ambassadress* and *sculpturer/sculptress*. This may give the illusion that English marks nouns for gender. This is not true, however, firstly because the suffix is not generically used with all nouns (we cannot say **directoress* or **doctoress*), and secondly because there is a great tendency to use the unmarked form of the noun with both sexes. Therefore the suffix is considered as "a matter of derivation, but not of grammatical gender."[2]

Since gender is eliminated in English, it does not mark adjectives according to the gender of the noun they modify. Yet the adjective *blond* is usually used with an additional ending *-e* when it is used to modify a feminine noun. This, however, does not break the rule, firstly because this is not a generic process of all English adjectives; secondly, the word's origin is French, a language which marks adjectives according to gender; and thirdly, the unmarked form can be used freely for both male and female nouns.

The pronoun system in English reflects the distinction between male, female, and sexless objects. Plamer divides nouns according to the way they can be referred to into seven types as shown in the following table[3]:

[1] C. E. Eckersley and J. M. Eckersley, *A Comprehensive English Grammar*, Essex: Longman, 1960, p. 42.
[2] Frank Palmer, *Grammar*, Middlesex: Penguin Books, 1971, p. 189.
[3] Ibid., pp. 189-190.

No.	Possible Sex(s)	Examples	Referring Pronoun(s)
1	male only	man, boy, king	he
2	female only	woman, girl, queen	she
3	neuter only	rock, chair, tree	it
4	male or female	student, teacher, player	he/she
5	male or neuter	stallion, bull, ram	he/it
6	female or neuter	mare, cow, ewe	she/it
7	male, female, or neuter	horse, sheep, cat	he/she/it

Table 17. English gender distribution

Some words such as *sun* may be referred to as *he*, *moon* referred to as *she*; mechanical things such as *ship*, *plane*, and *hovercraft* are referred to as *she*; names of countries such as *Egypt, France, Germany*, and *Britain* are usually referred to as *she*. This, however, does not indicate that the English nouns are marked for gender. First, these categories can indifferently be referred to as *it*. Second, "these are very few in number (and we should not wish to build a grammatical category on a few examples)."[1]

4.2.1.2. Number

English number has two features: singular and plural. In contrast with the Arabic number system, English does not make a special designation for dual. The default unmarked form of English nouns is the singular. The plural is formed either by the suffixation of an *-s* morpheme (such as *boy/boys, wish/wishes, study/studies,* and *wife/wives*); or by changing the word form (such as *man/men, mouse/mice*, and *thesis/theses*); or even without any change at all (*sheep/sheep*). According to Palmer[2], there are four conditions for regular plural nouns:

1. They have singular forms: *cats/cat, students/student,* and *children/child*.
2. They are used with plural verbs: *The students have arrived*.
3. They are used with numerals: *three children*.
4. They are used with plural demonstratives: *these cats*.

[1] Ibid., p. 190.
[2] Ibid., pp. 191-192.

Some categories of nouns are exceptionally considered plural while they do not meet the full plural requirements:

- Some nouns are always in the plural form and have no singular form, despite the fact that they indicate a single object, such as *scissors*, *shears* and *trousers*. These nouns are "semantically singular but syntactically plural."[1] These words occur with plural verbs and plural demonstratives: *These trousers are long.* But they cannot be used with numerals: * *three scissors*. Instead, we have to say: *three pairs of scissors.*
- The word *police* is an anomalous plural noun. It has no singular form. It takes a plural verb: *"The police have caught the murderer.*[2]*"* Yet the word does not occur with numerals: * *three police*, and does not occur with plural demonstratives: * *these police*.
- The words *people* and *cattle* have no singular form, but they function as plurals in all other respects: *Most people like traveling; three people; these people.*

Singular nouns can also be defined by explaining the environment in which they occur. There are four conditions for singular nouns:

1. They have plural forms: *cat/cats*, *student/students*, and *child/children*.
2. They are used with singular verbs: *The student has arrived*.
3. They can be used with indefinite articles *a*, *an* or the word *one*: *a child*; *an egg; one girl*. Moreover, they cannot occur without an article, whether definite or indefinite: *A boy went away*, or: *The boy went away*, but not: * *Boy went away*.
4. They are used with singular demonstratives: *this cat*.

Some categories of nouns are considered singular but do not meet the full singular requirements:

- The category of uncountable nouns (such as *bread*, *butter*, and *information*) has no plural form. They take singular verbs, but do not occur with the indefinite article *a* or *an*: * *a bread*. They can also occur without any article: *Butter is imported from abroad*.

---

[1] Alan Munn, "First Conjunct Agreement: Against a Clausal Analysis", *Linguistic Inquiry*, 30.4, 1999, 643-669, p. 645.
[2] *Longman Dictionary of English Language and Culture*, Essex, England: Longman, 1992, p. 1017.

- The category of collective nouns includes words such as *team*, *committee*, and *group*. These nouns may be used with a plural verb: *The team have won*, or with a singular verb: *The team has won*. They, however, must be referred to with a singular demonstrative pronoun: *this team*.

4.2.1.3. Case

English case has five features[1]: nominative, accusative, dative, genitive, and vocative. Nominative is the case of the subject of a finite verb such as *the boy* in *The boy smiles*. Accusative is the case of the direct object such as *football* in *He plays football* or the object of preposition such as *school* in *He went to school*. Dative is the case of the indirect object such as *John* in *I gave the book to John*. Genitive is the case denoting the possessor or owner such as *Jane's* in *Jane's book*. Vocative is the case denoting the person being called, addressed such as *boy* in *Come in, boy*.

Case is supposed to make a change of form of the words. This is true with pronouns, but with other nouns it is only the genitive case which is indicated by a word-ending, while "the other cases have lost their case-endings, and are indicated only by grammatical relation."[2]

4.2.1.4. Person

English person has three features: first, second, and third person. English uses seven lexical items to cover 18 places.

[1] C. E. Eckersley and J. M. Eckersley, *A Comprehensive English Grammar*, Essex: Longman, 1960, p. 45.
[2] J. C. Nesfield and F. T. Wood, *Manual of English Grammar and Composition*, London: Macmillian, 1964, p. 30.

Person	Gender	Singular	Plural
First	Masculine	I	We
	Feminine		
	Neuter		
Second	Masculine	You	
	Feminine		
	Neuter		
Third	Masculine	He	They
	Feminine	She	
	Neuter	It	

Table 18. English pronoun system

On Ingram's scale of personal pronouns, English is represented as a five-person system[1], with the following structure:

I	we
thou	
he	they

Table 19. English pronoun system according to Ingram

This means that English has no dual number. The pronouns *thou* (the English word is *you*) is not marked for number as it can be used both for the singular and the plural.

4.2.2. Agreement Degrees

Now I will classify agreement in English according to the degree of agreement required as follows:

4.2.2.1. Agreement in Number, Gender, and Person

Only two types of anaphors require this level of agreement: reflexive pronouns and possessive anaphors (or possessive adjectives).

[1] David Ingram, "Personal Pronouns", in Joseph H. Greenberg, ed., *Universals of Human Language*, Stanford, California: Stanford University Press, 1978, p. 243.

a. Reflexive Pronoun-Antecedent Agreement

The reflexive pronoun must agree with its antecedent in number: whether singular or plural; gender: whether feminine, masculine, or neuter; and person: whether first, second, or third person.

I_i depend on $myself_i$.
You_i depend on $yourself_i$.
You_i depend on $yourselves_i$.
He_i depends on $himself_i$.
She_i depends on $herself_i$.
It_i depends on $itself_i$.
We_i depend on $ourselves_i$.
$They_i$ depend on $themselves_i$.

In the above examples the reflexive pronouns and their antecedent pronouns are co-indexed to indicate that they co-refer to the same object.

b. Possessive Adjective-Antecedent Agreement

Possessive pronominals agree with their antecedents in number, gender, and person.

I_i depend on my_i father.
You_i depend on $your_i$ father.
He_i depends on his_i father.
She_i depends on her_i father.
It_i depends on its_i father.
We_i depend on our_i father.
$They_i$ depend on $their_i$ father.

The difference between possessive pronominals and reflexive pronouns is that reflexives must have their antecedent in the same domain and cannot have external reference. By contrast possessive pronominals can have external antecedents.

He_i depends on her_y father.
She_i depends on his_y father.

We notice in the above examples that the indexes of the pronouns and possessive pronouns do not match. This is because they do not co-refer to the same person.

4.2.2.2. Agreement in Number and Person

This degree of agreement is realized in English only in verb-subject agreement. English verb agrees with the subject in number in a very restrictive case: when the subject is a third person singular and the tense is the present. This is a highly selective case of agreement, which led many grammarians to question the sensitivity of the English verb for number. Palmer noted that "the present tense forms of the verb are not simply divided morphologically into singular and plural. The division is simply between the 'third person singular' and the rest".[1]

I play.	We play.
You play.	You play.
He plays.	They play.
She plays.	They play.
It plays.	They play.

In the past tense, however, even this slight agreement requirement disappears.

I played.	We played.
You played.	You played.
He played.	They played.
She played.	They played.
It played.	They played.

Verb *to be* is the only English verb that still retains a comparatively larger capability for agreement in number and person in the present and the past tenses.

Present	
I am happy.	We are happy.
You are happy.	You are happy.
He is happy.	They are happy.
She is happy.	They are happy.
It is happy.	They are happy.
Past	
I was happy.	We were happy.
You were happy.	You were happy.
He was happy.	They were happy.
She was happy.	They were happy.
It was happy.	They were happy.

[1] Frank Palmer, *Grammar*, Middlesex: Penguin Books, 1971, p. 191.

4.2.2.3. Agreement in Gender Only

Relative pronouns-antecedent agreement roughly makes the distinction between human and non-human. It does not make a strong gender distinction based on the two main categories of masculine and feminine. However, because neuter is included among the gender classification of the English noun, I considered the relative pronoun to have gender distinction. There is also agreement in number between the relative pronoun and the antecedent noun. This agreement is not explicitly shown by a change of the word form, but is implicitly indicated by the word behavior. The difference between *who plays* in the first sentence and *who play* in the second is a clear evidence that the relative pronoun has a number. However, because relative pronouns do not make overt representation of number, I excluded them from requiring number agreement.

The boy who plays in the garden is happy.
The boys who play in the garden are happy.

The girl who plays in the garden is happy.
The girls who play in the garden are happy.

The cat which plays in the garden is happy.
The cats which play in the garden are happy.

There is also alternative choices between *who* and *whom* according to their case whether nominative or accusative, respectively.

The boy who plays
The boy whom I met

4.2.2.4. Agreement in Number Only

a. Indefinite Article-Noun Agreement

Indefinite articles agree with the nouns they modify in number. The article *a* or *an* is used with the singular noun and zero article is used with the plural.

a boy
boys

With the definite article no agreement is required:

the boy
the boys

b. Demonstrative-Noun Agreement

Demonstrative adjectives agree with the nouns they modify in number. The demonstrative *this* and *that* are used with singular nouns; *these* and *those* are used with plural nouns.

this boy
these boys

that boy
those boys

c. Predicate-Subject Agreement

A predicate nominal (NP used as a complement of a copula verb) agrees in number with the subject. This can be illustrated by the following examples:

He is a student.
* He is students.

They are students.
* They are a student.

However, in some instances we can find singular predicate nominals while their subjects are plural and the sentences are still grammatical.

They are a problem.
They are the reason behind our success.

The solution to this problem can be found in the semantic restrictions imposed on the relationship between the subject and the predicate. These restrictions can be summarized as follows:

1. If the subject is singular the nominal predicate must be singular in all cases:

   This is a doctor.
   * This is doctors.

   This is a problem.
   * This is problems.

2. If the subject is plural and the predicate is a noun that denotes a physical object, the predicate nominal must be plural.

   These are doctors.
   * These are a doctor.

   These are books.
   * These are a book.

3. If the subject is plural and the predicate is a noun that denotes an abstract notion, the predicate nominal may be singular or plural.

These differences are an obstacle to peace.
These differences are obstacles to peace.

These factors are the reason behind our success.
These factors are the reasons behind our success.

These books are a gift.
These books are gifts.

4.2.2.5. Agreement in Case Only

This degree is found only in coordinated pronoun-antecedent agreement. When a pronoun is joined to a noun by a coordinating conjunction, "it should have the same case as its antecedent."[1]

(a) **They** and **I** are teachers.
(b) They will invite **him** or **her**.
(c) They gave it to **him** and **me**.
(d) **My** and **his** cars are white.

In example (a) the pronoun *I* agrees in case with *they*. They are both in the nominative case as they are in the subject position. In example (b) the two pronouns are in the accusative case; in example (c) they are in the dative case; and in example (d) the pronouns are both in the genitive case.

4.3. Cross-Language Redundancies

There is a considerable contrast between English and Arabic regarding agreement features. Whereas Arabic has strict rules for agreement between many parts of the sentence, English shows comparatively little consideration for these features. This makes mechanical translation between the two languages relatively cumbersome. When translating from Arabic into English the sentences will carry much agreement information which are redundant in English and therefore will be discarded. This poses no problems. The difficulty, however, appears when translating from English into Arabic, in which case much agreement information is not transferred.

[1] Maxine Hairston, *Contemporary Composition*, 4th ed., Boston: Houghton Mifflin Company, 1986, p. 537.

We can draw a map between the two languages to see how much agreement information each language manipulates compared to the other language:

Serial	Agreement Features	Arabic	English
1	Adjective-noun agreement in number	✓	×
2	Adjective-noun agreement in gender	✓	×
3	Adjective-noun agreement in person	✓	×
4	Adjective-noun agreement in case	✓	×
5	Adjective-noun agreement in definiteness	✓	×
6	Verb-Subject agreement in number	✓	✓
7	Verb-Subject agreement in gender	✓	×
8	Verb-Subject agreement in person	✓	✓
9	Demonstrative-noun agreement in number	✓	✓
10	Demonstrative-noun agreement in gender	✓	×
11	Demonstrative-noun agreement in case	✓	×
12	Relative pronoun-antecedent agreement in number	✓	×
13	Relative pronoun-antecedent agreement in gender	✓	✓
14	Relative pronoun-antecedent agreement in case	✓	✓
15	Anaphora[1]-antecedent agreement in number	✓	✓
16	Anaphora-antecedent agreement in gender	✓	✓
17	Anaphora-antecedent agreement in person	✓	✓
18	Predicative adjective-subject agreement in number	✓	×
19	Predicative adjective-subject agreement in gender	✓	×
20	Predicative nominal-subject agreement in number	✓	✓
21	Predicative nominal-subject agreement in gender	✓	×
22	Indefinite article-noun agreement in number	×	✓
23	Targets with double controllers	✓	×
24	Anti-agreement in number	✓	×
25	Anti-agreement in gender	✓	×
Score		24	10

Table 20. Agreement comparison between Arabic and English

[1] By anaphora here I mean only reflexives and possessives.

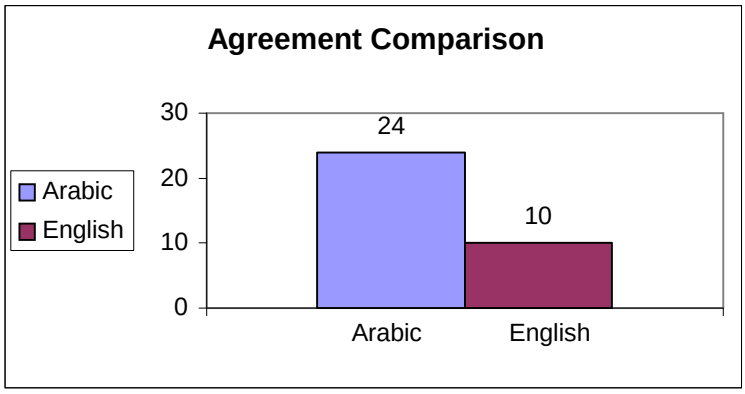


Fig. 20. Agreement comparison between Arabic and English

The above statistics show that English scored 10 agreement features while Arabic scored 24. Although these scores show how little agreement English employs in contrast with Arabic, this is not even expressive of the true facts. In cases where English scores positively for agreement, it does not employ the same morphological varieties which Arabic utilizes. The Arabic number system has a dual, which is not found in English. Arabic verbs, whether in the present or past, active or passive, agree with their subjects. English verb-subject agreement, contrastively, is extremely limited (third person singular in the present tense). Arabic also has a larger variety of personal pronouns. In short, Arabic employs agreement almost to the maximum, whereas English merely touches on the features.

Chapter Five
The Role of Agreement in the Translation Process

Agreement plays a considerable role in all phases of the translation process whether in analysis, transfer, or generation. In this chapter I will explore the role played by agreement in these phases.

5.1. The Role of Agreement in Analysis

The lexical database along with the morphological component collaborate during the analysis phase to define the number and gender of nouns in a sentence. The lexical database, as mentioned previously, includes information on grammatical categories, such as feminine, masculine, singular, and plural. And the morphological component provided with the system can analyze and interpret the different morphemes which may be attached to a word and which indicate number and gender with nouns and tense with verbs. Irregular forms of inflection are stored in a separate list, which is referenced by the system in due course.

The first step in parsing a sentence is to recognize the grammatical category of each word before deciding which group of words make a constituent phrase in the sentence. "Morphological analysis alone is sometimes sufficient to identify grammatical categories and structural functions, e.g. in English the suffix *-ize* usually indicates a verb."[1] English, however, is not rich in morphological inflection which indicates subject-verb and adjective-noun agreement. Even the suffix *-es* which indicates plural in nouns is also the same as that indicating agreement in present tense second person. Let us look at the following sentence:

Talk shows increase time waste.

The morphology alone cannot tell us the grammatical category of each word in the above sentence. Each word can function both as a noun and as a verb. We need to look at the grammatical context and grammatical relation between words to decide their categories. In this way "morphological analysis cannot be divorced from syntactic analysis."[2] To

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 83.
[2] Ibid.

parse the above sentence, we do not only need a phrase structure rule which states that:

S → NP VP

but we also need to introduce feature notations on this rule to explain the relationship between constituent parts. The rule must specify that the first NP in S must agree in number with the VP. The rule could be written as follows[1]:

1. S → NP[num=$X] VP[num=$X]
2. VP[num=$Y] → v[num=$Y] NP[num=?]
3. NP[num=$Z] → det (adj) n[num=$Z]
4. NP[num=$Z] → pron[num=$Z]

The first rule specifies that the NP must have the same number as the VP. The sign '$' preceding X, Y and Z means that it is a variable as distinguished from real values. The second rule states that when the VP is assigned a number it is the verb that carries the number specifications; the number of the second NP is not relevant. The third rule states that when an NP is assigned a number, it is the head noun that carries the number specifications. The fourth rule states that when the NP is a pronoun, the pronoun carries the number specifications.

The above rules can be used in a programming language to parse a sentence and check its validity by a simple condition:

s($A,$B)
if $A=np(…) *and* $B=vp(…) *and* number($A)=number($B)[2]

The rule states that the sentence is composed of two variables (a variable can stand for a word or group of words). If the first part meets the condition of an NP and the second part meets the conditions of a VP and the number of the first part equals the number of the second part, then the sentence is valid.

Applying these rules helps the system in identifying that the word *increase* in the above sentence is the verb of the sentence and hence helps

[1] Ibid., p. 60.
[2] Ibid., p. 66.

in demarcating different constituents of the sentence. The above rule explains the grammatical implausibility of sentences such as:

* The boy play football.
* The boys plays football.

The system can find that the number of *boy* in the first sentence is singular and the verb *play* is plural and the opposite with the second sentence. They do not meet the agreement requirement because they do not agree in number. They can, therefore, be judged by the system as ungrammatical sentences.

The above examples show how agreement, despite its limited scope of application in English, helps in the parsing process in both identifying grammatical categories of words and constituent parts of a sentence.

A pronoun agrees with its antecedent in number, gender, and person. This relationship can be made explicit in a parse tree by the introduction of either a pointer leading from one branch to the other or by the notion of co-indexing. Co-indexing means the attachment of an arbitrary number to an element and giving the same number to the other element which is linked to it. Both techniques are shown in the following figure:

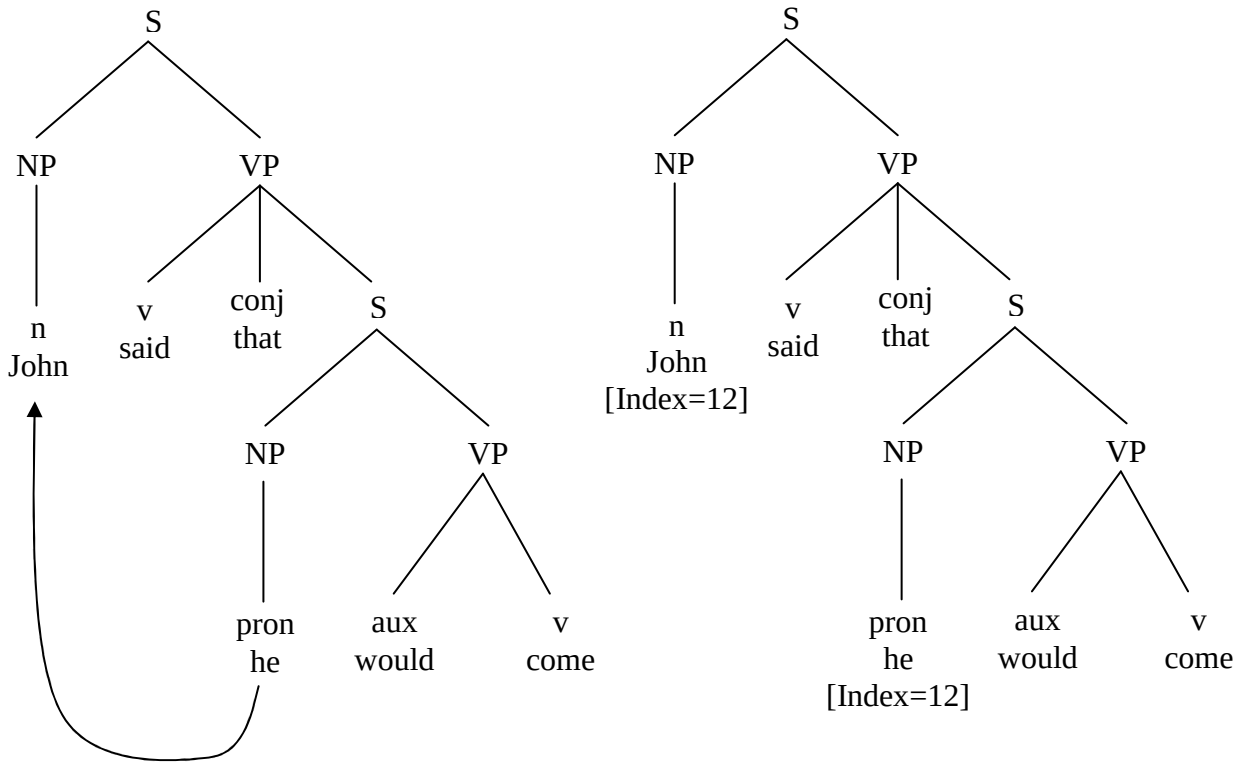


Fig. 21. Trees labeled with pointers and co-indexes[1]

The attachment of an index number to the noun and the pronoun that refers to it is a preferred technique in computational terms.

Our analysis, however, should ensure that the pronoun agrees in number and gender with its antecedent noun. So, the tree must not only carry information on the constituent phrases and grammatical categories of words, but it must also carry information on the syntactic and semantic functions of elements. Moreover, by applying complete morphological analysis, words are represented in their base forms. For a sentence such as: *This young girl likes her new bag*, the complete analysis will be as follows:

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, 1992, p. 62.

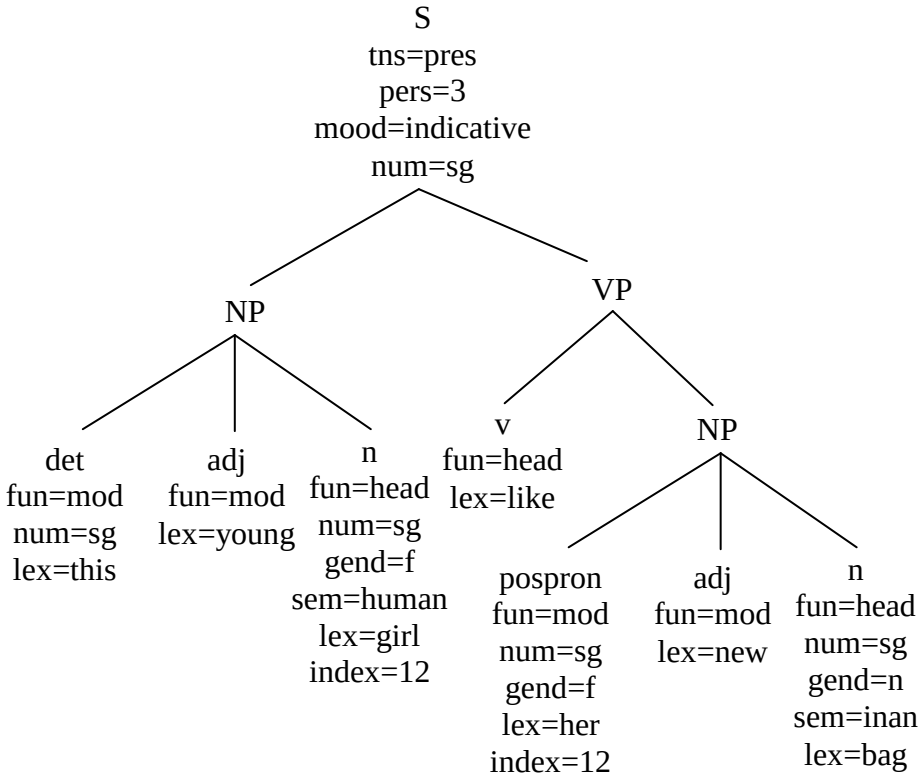


Fig. 22. A parse tree with full semantic and syntactic features

The above tree ensures that the anaphoric pronoun has the same number and gender as the antecedent and so it is co-indexed with it. We notice that "by matching up values for 'sex' and 'num', it is possible to recognize the anaphoric relation"[1] between pronouns and antecedents as well as other relations which require agreement. This relation can be either established or excluded. If, for example, the number or gender is different this means that the antecedent is not mentioned in the sentence and the co-indexing cannot be established with the noun. Moreover, the tree shows agreement in number between the demonstrative pronoun and the noun it modifies.

5.2. The Role of Agreement in Transfer

During lexical transfer, words (which have been morphologically, syntactically, and semantically analyzed) are looked up in the bilingual dictionary, and an equivalence map is drawn between source and target

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 111.

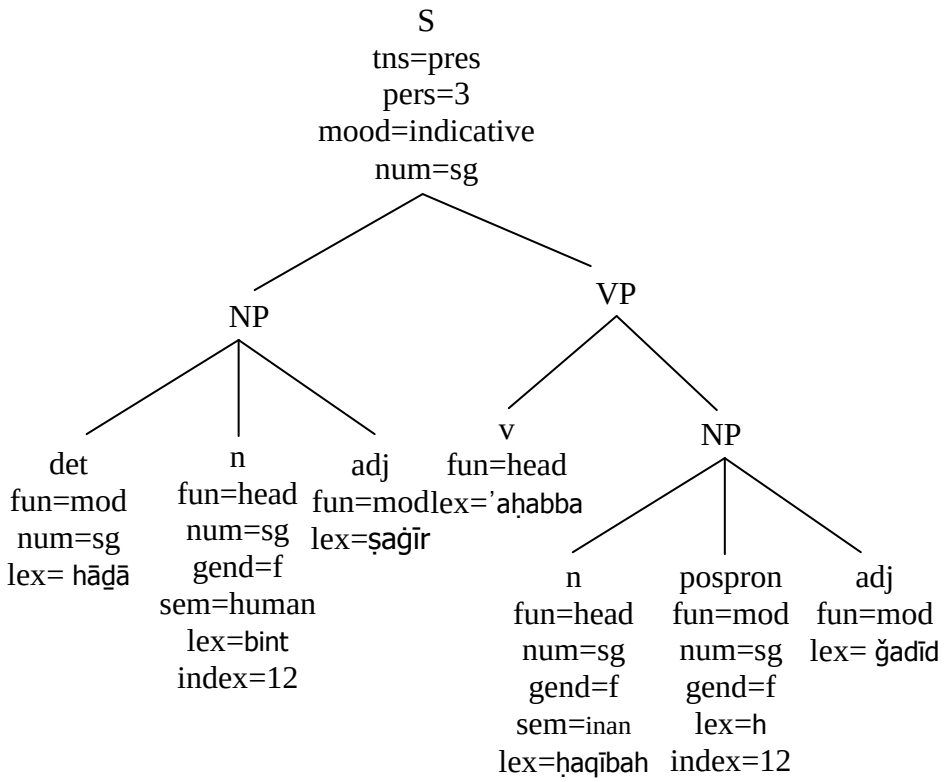
words. For our sample sentence, *This young girl likes her new bag*, the mapping will be as follows:

this *cat=det* *fun=mod* *num-sg*	⇒	hāḏā *cat=det* *fun=mod* *num-sg*
young *cat=adj* *fun=mod*	⇒	ṣaġīr *cat=adj* *fun=mod*
girl *cat=n* *fun=head* *num=sg* *sex=f* *sem=human*	⇒	bint *cat=n* *fun=head* *num=sg* *sex=f* *sem=human*
like *cat=v* *fun=head* *tns=pres*	⇒	ʾaḥabba *cat=v* *fun=head* *tns=pres*
her *cat=pospron* *fun=mod* *num=sg* *sex=f*	⇒	h *cat=pospron* *fun=mod* *num=sg* *sex=f*
new *cat=adj* *fun=mod*	⇒	ǧadīd *cat=adj* *fun=mod*
bag *cat=n* *fun=head* *num=sg* *sex=n* *sem=inan*	⇒	ḥaqībah *cat=n* *fun=head* *num=sg* *gender=f* *sem=inan*

Table 21. Lexical transfer

While most of the grammatical and semantic features are copied from the source language to the target language, we notice that the gender of nouns is determined by the gender specifications in the target language. This is shown by the word *bag* which is neuter in English, but its equivalent in Arabic, *ḥaqībah*, is feminine.

During structural transfer, tree-to-tree transfer carries all required features. The above parse tree of the English sentence, *This young girl likes her new bag*, will be rendered in Arabic as follows:



A Parse Tree after Transfer into Arabic

The structural transfer made the required changes to ensure that the tree structure suits the Arabic syntactic structure. For example, adjectives are moved to be after the nouns they modify and the possessive pronoun's position is also moved. However, the anaphoric relation is maintained and the co-index is transferred unchanged.

5.3. The Role of Agreement in Generation

The output of the transfer phase is a deep structure representation of the sentence with labeled nodes and target language lexical items in their base form. Of course the syntactic transfer made changes to make the word order suit the target language, but still some lexical changes and many morphological changes are required to produce an acceptable target language sentence.

During the process known as syntactic generation, the main task is "to order constituents in the correct sequence for the target language."[1] This reordering covers issues that were not dealt with during syntactic transfer such as the order of constituents when the sentence is labeled as 'passive' or 'interrogative'. In this case syntactic generation makes movements and/or adds words to meet the required feature.

A major task of the syntactic generation module is the "distribution of number and gender information to relevant terminal nodes."[2] Moreover, it assigns new number and gender features according to the requirements of the agreement rules in the target language. After the interaction of the syntactic generation module with our parse tree for the sentence, *This young girl likes her new bag*, it will look as follows.

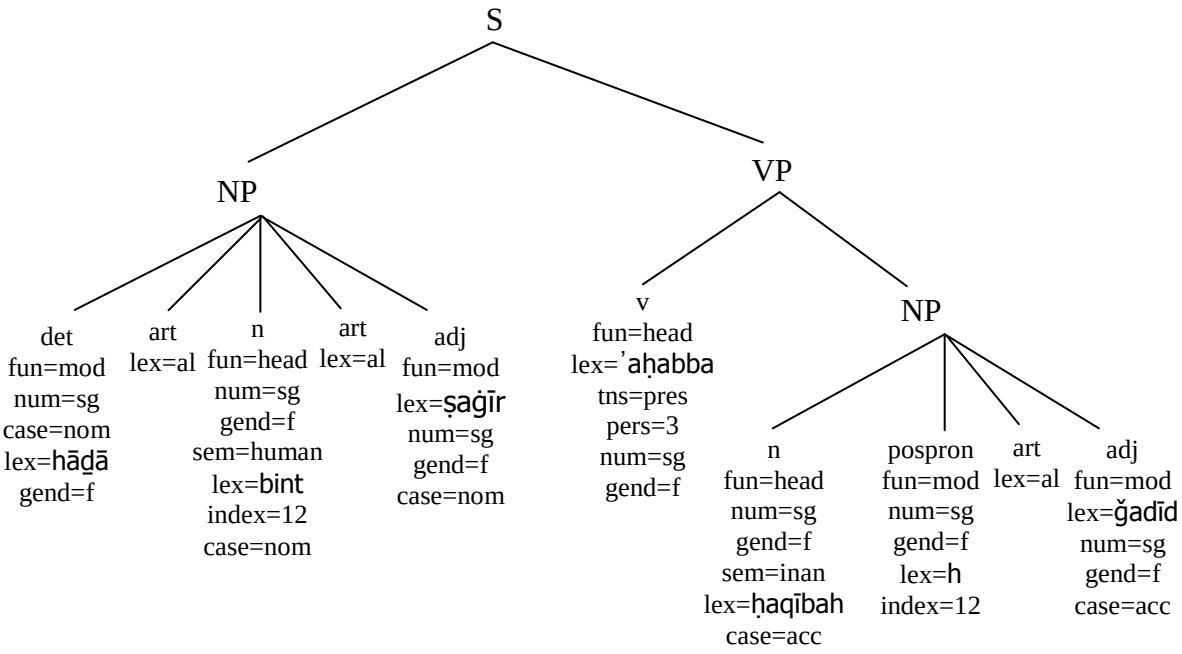


Fig. 23. A parse tree after syntactic generation

We notice here that the features of tense, person and number are moved from the sentence node to the relevant terminal node: the verb. Besides, in compliance with the requirement of the agreement rules, the syntactic generation component made new assignments of lexical features as follows:

[1] Ibid., p. 133.
[2] Ibid., p. 134.

1. The verb is assigned a gender because verbs in Arabic are required to agree with the subjects in gender as well as person and number.
2. The determiner is assigned a gender and case because demonstratives in Arabic are required to agree with the nouns they modify in gender and case as well as number.
3. The adjectives are assigned number, gender, and case because adjectives in Arabic are required to agree with the noun they modify in number, gender, and case.
4. Also in compliance with the requirement of the agreement rules, the syntactic generation component added the definite article *ʾal* before adjectives which modify definite nouns (*ʾal-ṣaġīr*) and before nouns which are preceded by demonstrative pronouns (*ʾal-bint*).

The last phase of generation and of the whole translation process is the morphological generation. It is a straightforward process which "interprets strings of labeled lexical items for output as target sentences."[1] Each time the morphological generator meets a terminal node with labels, it transforms it into actual morphological realization according to the morphological rules which handle general and special cases.

1. hāḏā [nbr=sg, gend=f, case=nom] = hāḏihi
2. bint [nbr=sg, gend=f, case=nom] = bintu
3. ṣaġīr[nbr=sg, gend=f, case=nom] = saġīratu
4. aḥabba [tns=pres, pers=3, nbr=sg, gend=f] = tuḥibbu
5. ḥaqībah [nbr=sg, gend=f, case=acc] = ḥaqībata
6. h [nbr=sg, gend=f] = ha
7. ǧadīd [nbr=sg, gend=f, case=acc] = ǧadīdata

The morphological generator is also responsible for combining some elements into a single word. For example, the definite article - *ʾal* is morphologically attached to the beginning of words (as a prefix) and possessive pronouns are attached to the end of words (as a suffix). The Arabic grammar states that the noun modified by a demonstrative pronoun is prefixed with - *ʾal*. After applying these last two rules, we will have the syntactically well-formed, grammatically accepted sentence:

hāḏihi-l-bintu-ṣ-ṣaġīratu tuḥibbu ḥaqībataha-l-ǧadīdata.

[1] Ibid., p. 133.

Chapter Six
Agreement Problems in MT

In this section I will explore different areas that cause (or are expected to cause) agreement problems during translation from English into Arabic. To show how a certain grammatical phenomenon causes an agreement problem, I will make use of 'test examples'. These examples are made-up sentences or phrases that contain an anticipated problem. The test examples will be put to the MT system (Al-Mutarjim Al-Arabey by ATA Software[1]) to see how it will deal with them. The performance will be analyzed in the 'test analysis' section.

All test examples in this thesis are collected and added in Appendix I to constitute an agreement 'test suite' for English-Arabic translation. A test suite is defined as a collection of "specially constructed test sentences. Each sentence in the suite contains either one linguistic construction of interest or a combination thereof."[2]

The difference between a test suite and a corpus should be emphasized. A corpus is a large collection of naturally occurring writings. A corpus is constructed by collecting sentences from books, encyclopedias, newspapers, magazines, radio, TV etc. A test suite, however, is a collection of made-up sentences. These sentences are constructed by an MT developer or specialist to explore different linguistic phenomena and grammatical constructions. The advantage of test suites over corpora is that test suites are more focused, shorter, and easier to use. Moreover, a test suite has more ability to reveal problems than a corpus. With a corpus "many potential areas of difficulty are hidden because the statistics are such that even quite large corpora will lack even a single example of particular grammatical combinations of linguistic phenomena."[3]

Test suites are very important in MT development and enhancement. Rules in an MT system may be "too strict or too general to apply correctly in all circumstances"[4]. Test suites help developers to detect such errors easily. They can then modify the rules and run the test suite again to measure improvements.

---

[1] Demo CD attached in Appendix II.
[2] Doug Arnold, et al., *Machine Translation: An Introductory Guide*, Manchester: Blackwell, 1994, p. 176.
[3] Ibid.
[4] Ibid., p. 175.

I would like to mention some facts about the test suite I devised in this thesis. First, It is useful only in testing agreement errors. It is not meant to measure translation performance in general. Second, it is designed for English-Arabic translation. It explores the agreement specifications and requirements in this language pair and in this specific translation direction. Third, my aim with the test examples is not to evaluate the performance of any MT system or compare it with others. My sole aim is to explore agreement problematic areas. Fourth, in my analysis of the test examples, I will focus only on agreement. The translation may contain errors in grammar, spelling, or morphological generation. I will not point out or analyze these errors in any way. The intention is not to measure the performance in general, but is confined to exploring the agreement phenomena and the problems they might cause in different grammatical constructions.

The generation component in an MT system is the module that is responsible for generating correct output sentences. It is fed with the necessary rules to make acceptably grammatical sentences in the output. Among the rules of Arabic generation are the agreement rules which specify what constituents require agreement, what their controllers are, and how many agreement features are involved. An MT system is usually able to apply regular agreement rules in the output sentences. It can make adjective-noun agreement.

Test Examples: Adjective-Noun Agreement

a. A diligent rich handsome man
رجل وسيم غني مجتهد
raǧul wasīm ġanī muǧtahid

b. A diligent rich handsome woman
إمرأة وسيمة غنية مجتهدة
ʾimraʾah wasīmah ġaniyyah muǧtahidah

c. Diligent rich handsome men
رجال وسيمون أغنياء مجتهدون
riǧāl wasīmūn ʾaġniyaʾ muǧtahidūn

d. Diligent rich handsome women
نِساء وسيمات غنيات مجتهدات
nisāʾ wasīmāt ġaniyyāt muǧtahidāt

e. I saw the diligent rich handsome men.
رَأيتُ الرّجالَ الوسيمين الأغنياء المجتهدين.
raʾaitu-r-riǧāla-l-wasīmīa-l-ʾaġniyāʾa-l-muǧtahidīn

Test Analysis:

The system successfully makes the adjectives agree in number and gender with their head noun. Example *e* shows correct agreement in gender, definiteness, and case, as the three adjectives following the noun show. It is not known why the system puts diacritics on some words and not on others. However, it is noted that diacritics appear mostly with complete sentences.

The system is also expected to successfully make agreement between a relative pronoun and the preceding noun.

Test Examples: Relative Pronoun-Noun Agreement

a. The man who drives the car
الرّجل الذي يَسُوقُ السّيارة
ʾar-rağulu-l-laḏī yasūqu-s-sayyārah

b. The woman who drives the car
الإمرأة التي تَسُوقُ السّيارة
ʾal-ʾimraʾatu-l-latī tasūqu-s-sayyārah

c. The men who drive the car
الرّجال الذين يَسُوقونَ السّيارة
ʾar-riğālu-l-laḏīna yasūqūna-s-sayyārah

d. The women who drive the car
النِّساء اللواتي يَسُوقنَ السّيارة
ʾan-nisāʾu-l-lawātī yasūqna-s-sayyārah

Test Analysis:

In examples *a* through *d* the system correctly places the relative pronoun that agrees in number and gender. No test on agreement on case is possible because relative pronouns have overt case marking only in the dual, one of the issues that cause problems to MT as will be shown later.

The system can also make agreement between demonstratives and the nouns they modify.

Test Examples: Demonstrative Adjective-Noun Agreement

a. This man
هذا الرّجل
hāḏā-r-rağuli

b. This woman
هذه الإمرأةِ
hāḏihi-l-ʾimraʾati

c. These men
هؤلاء الرّجال
hāʾulāʾi-r-riğāli

d. These women
هؤلاء النِّساءِ
hāʾulāʾi-n-nisāʾi

e. That man
ذلك الرّجل
ḏālika-r-rağuli

f. That woman
تلك الإمرأةِ
tilka-l-ʾimraʾahi

g. Those men
أولئك الرّجال
ʾulāʾika-r-riǧāli

h. Those women
أولئك النّساء
ʾulāʾika-n-nisāʾi

Test Analysis:

The system correctly makes agreement in number and gender between demonstratives and their head nouns. No test on agreement in case is possible because demonstratives have overt case marking only in the dual, one of the issues that cause problems to MT as will be shown later.

The system can also make agreement between a verb and its subject.

Test Examples: Verb-Subject Agreement

a. The boy goes to the garden and waters the flowers.
يَذْهبُ الولدُ إلى الحديقةِ ويَسْقي الزّهور.
yaḏhabu-l-waladu ʾilā-l-ḥadīqati wa yasqī-z-zuhūr.

b. The girl goes to the garden and waters the flowers.
تَذْهبُ البنتُ إلى الحديقةِ وتَسْقي الزّهور.
taḏhabu-l-bintu ʾilā-l-ḥadīqati wa tasqī-z-zuhūr.

c. The boys go to the garden and water the flowers.
الأولاد يَذْهبونَ إلى الحديقةِ ويَسْقونَ الزّهور.
ʾal-ʾawlādu yaḏhabūna ʾilā-l-ḥadīqati wa yasqūna-z-zuhūr.

d. The girls go to the garden and water the flowers.
البنات يَذْهبنَ إلى الحديقةِ ويَسْقينَ الزّهور.
ʾal-banātu yaḏhabna ʾilā-l-ḥadīqati wa yasqīna-z-zuhūr.

e. The boy went to the garden and watered the flowers.
ذهبَ الولدُ إلى الحديقةِ وسَقى الزّهور.
ḏahaba-l-waladu ʾilā-l-hadīqati wa saqā-z-zuhūr.

f. The girl went to the garden and watered the flowers.
ذهبتْ البنتُ إلى الحديقةِ وسَقتْ الزّهور.
ḏahabati-l-bintu ʾilā-l-ḥadīqati wa saqati-z-zuhūr.

g. The boys went to the garden and watered the flowers.
ذهبوا الأولادُ إلى الحديقةِ وسَقوا الزّهور.
ḏahabū-l-ʾawlādu-ʾilā-l-ḥadīqati wa saqū-z-zuhūr.

h. The girls went to the garden and watered the flowers.
ذهبنَ البناتُ إلى الحديقةِ وسَقينَ الزّهور.
ḏahabna-l-banātu ʾilā-l-ḥadīqati wa saqaina-z-zuhūr.

Test Analysis:

The system correctly makes agreement between verb and subject in gender in examples *a* and *b* where the verb precedes the subject, and in gender,

number and person in examples *c* and *d* where the verb follows its subject. However, agreement in *g* and *h* is faulty, as the verb agrees in number with the subject while the verb precedes the subject. This is not permitted in Standard Arabic.

An MT system is also expected to follow the rules related to agreement and anti-agreement in gender and number between cardinal numbers and their head nouns.

Test Examples: Agreement in the Number System

a. one boy
ولد واحد
waladun wāḥid

b. one girl
بنت واحدة
bintun wāḥidah

c. two boys
ولدان
waladān

d. two girls
بنتان
bintān

e. three boys
ثلاثة أولاد
ṯalāṯatu ʾawlād

f. three girls
ثلاث بنات
ṯalāṯu banāt

g. four boys
أربعة أولاد
ʾarbaʿatu ʾawlād

h. four girls
أربع بنات
ʾarbaʿu banāt

i. ten boys
عشَر أولاد
ʿašru ʾawlād

j. ten girls
عشْرة بنات
ʿašratu banāt

k. eleven boys
أحدَ عشرَ ولد
ʾaḥada ʿašra walad

l. eleven girls
إحدى عشْرة بنت
ʾiḥdā ʿašrata bint

m. twelve boys
إثنا عشرَ ولد
ʾiṯnā ʿašra walad

n. twelve girls
إثنتا عشْرة بنت
ʾiṯnatā ʿašrata bint

o. fifteen boys
خمسة عشرَ ولد
ẖamsatu ʿašra walad

p. fifteen girls
خمس عشْرة بنت
ẖamsu ʿašrata bint

q. twenty-one boys
واحد وعشرون ولد
waḥidun wa ʿišrūna walad

r. twenty-one girls
واحدة وعشرون بنت
wāḥidatun wa ʿišrūna bint

s. seventy-eight boys
ثمانية وسبعون ولد
ṯamāniyatun wa sabʿūna walad

t. seventy-eight girls
ثمان وسبعون بنت
ṯamānin wa sabʿūna bint

Test Analysis:

The system successfully follows the rules for anti-agreement in number and gender in all examples except *i* and *j*. There is also a fault in case assignment for the head noun. All *walad* should be *waladan*, and all *bint* should be *bintan*.

However, agreement cannot always be realized as readily as seen above. This is due to the fact that Arabic differs greatly from English in the distribution of number and gender in the pronoun system, lexical items as well as the syntactic structure. This difference results in many agreement problems during the translation process. These problems will be investigated in detail in the following sections.

6.1. Pronouns

Only the pronouns *he* and *she* do not cause an agreement problem during translation into Arabic because they are clearly marked for number and gender. The other English pronouns *you*, *they*, *it*, *I* and *we* cause an agreement problem. This is due to the fact that the Arabic pronoun system differs from the English one in that the Arabic system includes a larger number of pronouns to allow for the distribution of features such as: singular, dual, plural, feminine, and masculine.

The pronoun *you* is not marked for gender and its number is ambiguous as it can refer to singular, dual, and plural entities. The translation of this pronoun as well as the agreement specifications of the target words depends on clear identification of these features:

You (m, sg) are kind. ʾanta ʿaṭūf	You (f, sg) are kind. ʾanti ʿaṭūfah
You (m, dl) are kind. ʾantumā ʿaṭūfān	You (f, dl) are kind. ʾantumā ʿaṭūfatān
You (m, pl) are kind. ʾantum ʿaṭūfūn	You (f, pl) are kind. ʾantunna ʿaṭūfāt

Test Examples: The Pronoun *You*

a. You are a good boy. أنتَ ولدُ جيدُ. ʾanta waladu ǧayyidu.	b. You are a good girl. أنتَ بنتُ جيدةُ. ʾanta bintu ǧayyidatu.
c. You are two good boys. أنتَ ولدان جيدون. ʾanta waladāni ǧayyidūna.	d. You are two good girls. أنتَ بنتان جيداتُ. ʾanta bintāni ǧayyidātu.
e. You are good boys.	f. You are good girls.

أنتَ بناتُ جيداتُ.
ʾanta banātu ǧayyidātu.

أنتَ أولادُ جيدون.
ʾanta ʾawlādu ǧayyidūna.

g. You are good.
أنتَ جيدُ.
ʾanta ǧayyidu.

Test Analysis:

The system uses the default masculine singular form of the pronoun in all cases. This is correct only in examples *a* and *g*. In all other examples, pronoun choice is clearly wrong.

Similarly, the English pronoun *they* is not marked for gender and its number is ambiguous from the Arabic point of view which distinguishes between dual and plural:

They (m, dl) are kind.
humā ʿaṭūfāni.

They (f, dl) are kind.
humā ʿaṭūfatāni.

They (m, pl) are kind.
hum ʿaṭūfūna.

They (f, pl) are kind.
hunna ʿaṭūfātin.

They (pl [non-human]) are kind.
hiya ʿaṭūfah

Test Examples: The Pronoun *They*

a. They are two good boys.
هم ولدان جيدون.
hum waladāni ǧayyidūn.

b. They are two good girls.
هم بنتان جيداتُ.
hum bintāni ǧayyidāt.

c. They are good boys.
هم أولادُ جيدون.
hum ʾawlādun ǧayyidūn.

d. They are good girls.
هم بناتُ جيداتُ.
hum banātun ǧayyidāt.

e. They are good cats.
هم قططُ جيدةُ.
hum qiṭaṭun ǧayyidah.

Test Analysis:

The system uses the default masculine plural form of the pronoun in all cases. This is correct only in example *b*. In all other examples, pronoun choice is wrong.

The English pronoun *it* is not marked for gender. It is not clear whether it refers to a masculine or feminine object. Arabic, however, needs this distinction.

It (m) is good.
huwa ğayyidun.

It (f) is good.
hiya ğayyidatun.

Test Examples: Pronoun *It*

a. It is a good bull.
إنه ثور جيد.
ʾinnahu ṯawrun ğayyidun.

b. It is a good cow.
إنه بقرة جيدة.
ʾinnahu baqaratun ğayyidatun.

Test Analysis:

The system uses the default masculine singular form of the pronoun in all cases. This is correct only in example *a*. In example *b* the pronoun choice does not agree in gender with the noun it refers to.

The pronoun *we* is not marked for gender and its number, from the perspective of Arabic, is ambiguous (dual or plural). Arabic does not have translational varieties of the pronoun, yet some agreement specifications depend on these features:

We (m, dl) are kind.
naḥnu ʿaṭūfani.

We (f, dl) are kind.
naḥnu ʿaṭūfatāni.

We (m, pl) are kind.
naḥnu ʿaṭūfūna.

We (f, pl) are kind.
naḥnu ʿaṭūfātun.

Test Examples: The Pronoun *We*

a. The two boys said, “We are good.”
إنّ الولدين قالا، " نحن جيدون. "
ʾinna-l-waladaini qālā, “naḥnu ğayyidūn”

b. The two girls said, “We are good.”
إنّ البنتين قالتَا، " نحن جيدون. "
ʾinna-l-bintaini qālatā, “naḥnu ğayyidūn”

c. The boys said, “We are good.”
إنّ الأولادَ قالوا، " نحن جيدون. "
ʾinna-l-ʾawlāda qālū, “naḥnu ğayyidūn”

d. The girls said, “We are good.”
إنّ البناتِ قلنَ، " نحن جيدون. "
ʾinna-l-banāti qulna, “naḥnu ğayyidūn”

Test Analysis:

Only the translation in example *c* is correct. Despite the fact that the pronoun *we* has only one translation, its hidden number and gender

specifications affect the choice of the form (gender and number) of the predicative adjectives.

The pronoun *I* is not marked for gender. Arabic does not have translational varieties of the pronoun, yet some agreement specifications depend on these features:

I (f) am kind.	I (f) am kind.
ʾanā ʿaṭūfun.	ʾanā ʿaṭūfatun.

Test Examples: Pronoun *I*

a. The boy said, “I am good.”
الولد قال، "أنا جيد."
ʾal-waladu qāla, “ʾanā ǧayyid.”

b. The girl said, “I am good.”
البنت قالت، "أنا جيد."
ʾal-bintu qālat, “ʾanā ǧayyid.”

Test Analysis:

Only the translation in example *a* is correct. Despite the fact that the pronoun *I* has only one translation, its hidden number and gender specifications affect the choice of the form (gender) of the predicative adjectives. It is not clear why the system does not use the emphatic article *ʾinna* here as in the previous test examples, and why the sentences here do not follow the preferred verb-subject order.

6.1.1. Proposed Solution

My proposed solution to this problem is to make the source language analyzer (or parser) assign the proper number and gender for each pronoun. As much as possible, the parser must specify whether the pronoun is masculine, feminine, singular, dual, or plural. This means that the parser will need to include a level of analysis which is specific to the target language. English does not normally mark some pronouns (like *you* and *they*) for gender. However, in anticipation of the target language requirements, the system will try to mark these pronouns for gender “since this information is relevant for translation”[1] into Arabic.

In trying to detect the gender and number of pronouns, the system can make use of linguistic clues like the reference of the pronoun. If the system succeeds in establishing the link between the pronoun and the

[1] Frank Van Eynde, “Machine Translation and Linguistic Motivation”, in Frank Van Eynde, ed., *Linguistic Issues in Machine Translation*, London: Pinter Publishers, 1993, p. 24.

object it refers to, then it can easily assign gender and number to the pronoun.

In case the system fails to recognize the correct gender and number of a pronoun, it can then resort to the best guess method which means deciding an interpretation "based on the relative likelihood of one analysis over the other, based only on which structures are more or less common"[1]. In this case the system will use the default values for gender and number.

6.2. Proper Nouns

In an MT system there is a module responsible for identifying proper names, which is usually referred to as "Proper-Name Recognizer"[2]. However, proper nouns[3] can cause confusion in MT in two different ways. The first, the MT system may not identify that the word in question is a proper noun and translates it as a common noun, adjective, or whatever can be found in the dictionary. The second is that it cannot identify the gender of the noun and thus fails to provide information needed to make agreement in Arabic.

The first case is the group of English proper nouns which have entries in main domain dictionaries and defined as common nouns, adjectives, or any other part of speech. They can be exemplified by the following names:

- Names of males: *Bush, Carpenter, Foot, Link, Lance, King, Ace, Chance, Rice, Black, Smith, Will, White, Bill, Knight,* and *Fox.*
- Names of females[4]: *Harmony, Holly, Honey, Hope, Flower, Flora, Aura, Bonnie, Carol, Cherry, Clemency, Lily, Bliss,* and *Blossom.*

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 94.
[2] Fred Popowich, et al., "Machine Translation of Closed Captions", *Machine Translation*, 15, 311-341, 2000, p. 322.
[3] Proper names and proper nouns are used synonymously.
[4] From Lareina Rule, *Name Your Baby*, New York: Bantam Books, 1963.

Test Examples: Meaning-Ambiguous Proper Nouns

a. Bush read the book.
قَرأتْ الأجمةُ الكتابَ.
qaraʾati-l-ağamatu-l-kitāb

b. Harmony read the book.
قَرأ الإنسجامُ الكتابَ.
qaraʾa-l-insiğāmu-l-kitāb

c. Carpenter read the book.
قَرأ النّجارُ الكتابَ.
qaraʾa-n-nağğāru-l-kitāb

d. Hope read the book.
قَرأ الأملُ الكتابَ.
qaraʾa-l-ʾamalu-l-kitāb

e. Foot read the book.
قَرأ القدمُ الكتابَ.
qaraʾa-l-qadamu-l-kitāb

f. Flora read the book.
قَرأتْ النباتاتُ الكتابَ.
qaraʾati-n-nabātātu-l-kitāb

g. Lance read the book.
قَرأ الرّمحُ الكتابَ.
qaraʾa-r-rumḥu-l-kitāb

h. Aura read the book.
الهالة قَرأتْ الكتابَ.
ʾal-hālatu qaraʾati-l-kitāb

Test Analysis:

The system fails to recognize any of the proper nouns above, and translates them as common nouns.

The second case is the group of proper nouns which can easily be identified as proper nouns since they have no entries in main domain dictionaries and have no other meanings. Yet the problem is to decide whether these names denote males or females. They can be exemplified by the following:

- English names of males: *Jack, Jackson, George, Addison, Adrian, Alexander, Alfred, Craig,* and *Oliver.*
- English names of females: *Jane, Janet, Jacqueline, Sarah, Nancy, Alice, Angela, Andrea, Anne,* and *Barbara.*

Moreover, the names in any translation passage do not need to be restricted to English names. A passage can include various names from various languages, thus adding to the complexities of identifying the sex of these names.

- French names of males: *François, Pierre, André, Gérard, Chirac, Bernhardt, Gounod, Barthélemy, Nicolas,* and *Antoine.*
- French names of females: *Françoise, Simone, Josephine, Rachel, Élisa, Catherine, Renée, Jeanne, Julia, Marie, Madeleine,* and *Marguerite.*
- Spanish names of males: *Vasco, Antonio, Juan, Rubén, José, Ribera, Benito, Manuel, Francisco, Jusepe, Alemán, Mateo,* and *Diego.*
- Spanish names of females: *Isabella, Maria, Marie, Christina, Cecilia,* and *Francisca.*

- Arabic names of males: *Ahmed, Hassan, Hashim, Kamal, Karim, Monir, Ali, Salim, Ramzi,* and *Fauzi.*
- Arabic names of females: *Salma, Nada, Karima, Kauthar, Hoda, Aziza, Marwa, Nora,* and *Samira.*

Test Examples: Gender-Ambiguous Proper Nouns

a. Jack read the book.
قَرأ جاك الكتاب.
qaraʾa ǧāku-l-kitāb

b. Jane read the book.
جان قَرأ الكتابَ.
ǧān qaraʾa-l-kitāb

c. Jackson read the book.
جاكسون قَرأ الكتابَ.
ǧāksūn qaraʾa-l-kitāb

d. Janet read the book.
جانيت قَرأ الكتابَ.
ǧānīt qaraʾa-l-kitāb

e. Nicolas read the book.
نيكولاس قَرأ الكتابَ.
nīkūlās qaraʾa-l-kitāb

f. Josephine read the book.
جوزيفين قَرأ الكتابَ.
ǧūzīfīn qaraʾa-l-kitāb

g. Pierre read the book.
بير قَرأ الكتابَ.
bīr qaraʾa-l-kitāb

h. Simone read the book.
سمون قَرأ الكتابَ.
simūn qaraʾa-l-kitāb

i. Vasco read the book.
فاسكو قَرأ الكتابَ.
fāskū qaraʾa-l-kitāb

j. Isabella read the book.
قَرأتْ إيزابيلا الكتاب.
qaraʾat ʿīzabīllā-l-kitāb

k. Antonio read the book.
أنتونيو قَرأ الكتابَ.
antūnyū qaraʾa-l-kitāb

l. Maria read the book.
ماريا قَرأ الكتابَ.
mariyyā qaraʾa-l-kitāb

m. Ahmad read the book.
قَرأ أحمد الكتاب.
qaraʾa ʾaḥmadu-l-kitāb

n. Salma read the book.
سالما قَرأ الكتابَ.
sālmā qaraʾa-l-kitāb

o. Hassan read the book.
قَرأ حسن الكتاب.
qaraʾa ḥasanu-l-kitāb

p. Nada read the book.
قَرأتْ ندى الكتاب.
qaraʾat nadā-l-kitāb

Test Analysis:

The system succeeds in recognizing the gender of the proper nouns of feminine names on the right column only in examples *j* and *p*. With the other examples it uses the default gender, i.e. masculine. Regarding masculine names on the left column, we are not sure whether the system recognizes them as masculine or merely uses the default gender. It is not clear why the system makes alternate word orders: SVO vs. VSO. Yet it can be noted that whenever the system uses VSO with feminine proper nouns, it correctly detects the gender.

6.2.1. Proposed Solutions

Firstly, to identify that the word is a proper noun and avoid confusing it with other categories, some solutions can be proposed:

1. Grammatical context: A proper noun must be always capitalized and not preceded by either a definite or indefinite article, whereas a common noun, when it is singular, must always be preceded by either a definite or indefinite article.
2. Semantic clues: The co-occurrence restrictions of the verb *reads*, for example, require a human entity in the subject position. This can help in the exclusion of the interpretation of *Flora*, for example, as a common noun.
3. Some titles which precede or follow names can help in identifying the proper noun, such as *PhD.*, *Prof.*, *Dr.*, *Senator, Governor, Chancellor, President, Prime Minister* and *Artist*.

Secondly, to decide the gender of the proper noun, we need to make use of some solutions:

1. Constructing a lexical database for proper nouns with gender specification. The lexical entry for each name will tell us whether the name is feminine or masculine. This solution, however, is not practical. First, some names may confusingly be used both for males and females. Second, it is hard to list all possible proper names in a language. Third, it is even harder to list in the lexical database all proper names in all languages.
2. Making use of titles which denote sex such as, *Mr.*, *Mrs.*, *Ms.*, *King, Queen, Duke, Duchess, Sir, Madame, Baron, Baroness, Father, Mother, Prince, Princess, Lord, Brother, Sister, Nurse* and *Rev.* If the name is preceded by a masculine title this definitely means that the name is masculine, and vice versa. These titles are frequently, but not always, used with names, and by manipulating them we can considerably reduce the ambiguity which results in agreement violation in the target language.
3. Resorting to the user interaction method. When the system fails to make an appropriate decision regarding the gender of a name either because it is from a foreign language or from the same language but not included in the system database, it may "ask human operators to select the analysis which conforms with their knowledge"[1]. User

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 94.

interaction can also be manipulated in another way. The system may provide the facility for the user to create his/her customized dictionary and store proper names in it. This facility is already provided by Al-Mutarjim Al-Arabey as shown in the following figure.

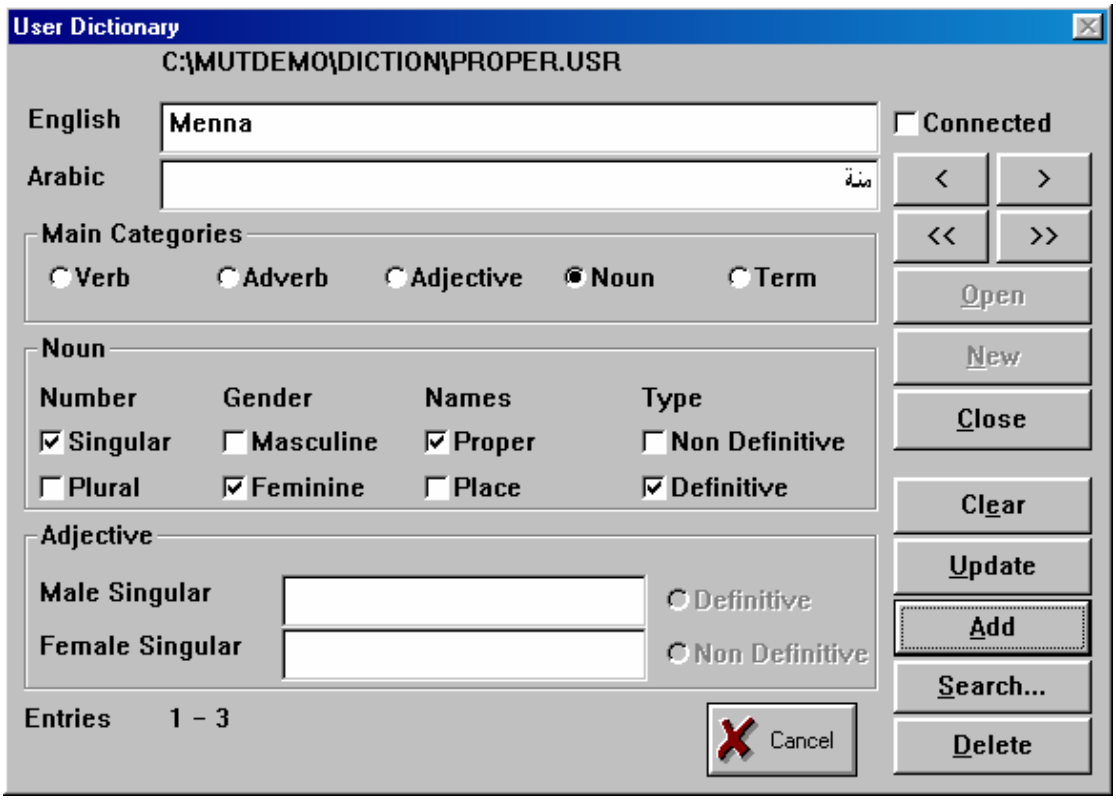


Fig. 24. Entering a new proper noun to the user dictionary

6.3. Common Nouns

English common nouns create a lexical transfer ambiguity which arises when "a single source language word can potentially be translated by a number of different target language words."[1] This ambiguity arises from the fact that almost all English nouns are gender-neutral, while Arabic nouns are gender-sensitive. In Arabic all inanimate nouns must be either masculine or feminine. This creates little or no problem in translation because the Arabic lexical entry will provide all information needed regarding the gender of the noun in question. The problem arises, however, from Arabic animate nouns which can have feminine and masculine variants, and, therefore, there is no one possible straightforward lexical substitution between English and Arabic.

[1] Ibid., p. 99.

English Word	Possible Arabic Equivalents
student	ṭālib (masculine) ṭālibah (feminine)
teacher	mudarris (masculine) mudarrisah (feminine)
player	lāʿib (masculine) lāʿibah (feminine)
seller	bāʾiʿ (masculine) bāʾiʿah (feminine)
engineer	muhandis (masculine) muhandisah (feminine)
cat	qiṭṭ (masculine) qiṭṭah (feminine)
dog	kalb (masculine) kalbah (feminine)

Table 22. Examples of lexically ambiguous English nouns

Although most translation systems choose the default unmarked masculine equivalent, this is not always a safe guess as the English noun may actually refer to a feminine entity. The only sure ground here is when the English noun explicitly refers either to a male or female; a category of words not so abundant in English:

English Nouns	Arabic Equivalent
king	malik
queen	malikah
hen	dağāğah
cock	dīk
prince	ʾamīr
princess	ʾamīrah

Table 23. Examples of non-ambiguous English nouns

Furthermore, English plural nouns are quadruple-folded ambiguous as the choice is now not only between masculine and feminine, but also between dual and plural.

English Word	Possible Arabic Equivalents
students	ṭālibān (dual masculine) ṭālibatān (dual feminine) ṭullāb (plural masculine) ṭālibāt (plural feminine)
cats	qiṭṭān (dual masculine) qiṭṭatān (dual feminine) qiṭaṭ (plural)

Table 24. Examples of lexically ambiguous English plural nouns

Test Examples: Common Nouns

a. She is a good singer.
هي مغنيُ جيدُ.
hiya muġannī ǧayyid

b. These girls are students.
هؤلاء البناتِ طلابُ.
hāʾulāʾi-l-banāti ṭullāb

c. These women are engineers.
هؤلاء النِّساءِ مهندسونُ.
hāʾulāʾi-n-nisāʾi muhandisūn

d. These women are teachers.
هؤلاء النِّساءِ معلمونُ.
hāʾulāʾi-n-nisāʾi muʿallimūn

e. These girls are good players.
هؤلاء البناتِ هن لاعبونُ جيداتُ.
hāʾulāʾi-l-banāti hunna lāʿibūna ǧayyidāt

f. These women are good actors.
هؤلاء النِّساءِ هن ممثلونُ جيداتُ.
hāʾulāʾi-n-nisāʾi hunna mumaṯilūna ǧayyidāt

g. The student likes her teacher.
الطّالب يَحْبُّ معلّمها.
aṭ-ṭalibu yuḥibbu muʿallimaha

h. The students like their teachers.
الطّلاب يَحْبّونَ معلميهم.
aṭ-ṭullābu yuḥibbūna muʿallimīhim

Test Analysis:

In examples *a* through *f* the predicative nominal is clearly feminine. English does not mark nouns or adjectives for gender, and so the system follows the English track and translates them as unmarked nouns, i.e. masculine, resulting in a clear violation of the agreement rules in Arabic. In example *g* the gender of *student* and *teacher* is ambiguous, yet the anaphoric pronominal *her* may propose that *student* is feminine. In example *h* both *students* and *teachers* are ambiguous and there is no clue to identify their gender, and so the system justifiably used the default gender. In examples *e* and *f* the system adds a pronoun (*hunna*) after the subject. It seems that the purpose is to fine-tune the output when the predicate is composed of a noun and an adjective. It is not clear in these two examples (*e* and *f*) why the adjectives agree with the subjects but not with the nouns they modify.

6.3.1. Proposed Solution

First, the Arabic lexical entry must include information on all possible interpretations of the common noun. It must make explicit the fact that the common noun can have masculine and feminine variants. This will make all options available for the generation modules to choose from. Second, the parser must provide, as much as possible, information on the gender of common nouns. Despite the fact that this information is not linguistically motivated by the source language, it is crucial in reaching a sound translation in the target language. This kind of analysis will be target-language specific, i.e. if the target language is not Arabic the system will not necessarily need to extract this information.

To decide on the gender of the noun, the parser must look for linguistic clues such as antecedents and referring pronouns. In many instances the sentence may be void of any clues to help in deciding on the gender of the noun. In these cases the system will have no other option but to use the default unmarked gender, that is the masculine.

6.4. The Dual

English has no dual. The English number system has only two meanings "ONE and OTHER/MORE THAN ONE."[1] In this way, English contrasts with Arabic, which can express the meanings of one, two, and more than two. This contrast poses a problem in translating "natural pairs such as eyes, hands and feet"[2]. If these natural pairs are rendered as plurals in a language that supports dual, they sound really odd as they speak about someone who has more than two hands or eyes, etc.

Test Examples: The Dual: Natural Pairs

a. I held him with my hands.
حَملته بأياديي.
ḥamaltuhu bi-ʾayādiyy

b. I saw him with my eyes.
رَأيته بعيوني.
raʾaituhu bi-ʿuyūnī

[1] Wallis Reid, *Verb and Noun Number in English: A Functional Explanation*, London: Longman, 1991, p. 122.
[2] Ibid.

c. I heard him with my ears.
سَمعته بآذاني.
samiʿtuhu bi-ʾāḏānī

d. My legs cannot carry me.
سيقاني لا يُمكنُ أنْ تَحْملني.
sīqānī la yumkinu ʾan taḥmilanī

Test Analysis:

The Arabic plural means "more than two", and so it sounds odd to talk about someone who has more than two hands, eyes, ears, or legs. The system should have interpreted English plural in the above examples as Arabic dual.

The dual can still be expressed in English in a number of ways: when the two nouns are joined by *and*, or when a plural noun is preceded by *two*, *couple of*, or *both*. The Arabic translation must be dual, requiring the demonstratives, relative pronouns, verbs, adjectives, etc. to agree with the noun in duality.

Test Examples: The Dual: Some Expressions

a. These two men
هؤلاء الرّجلين
hāʾulāʾi-r-rağulain

b. These two women
هؤلاء الإمرأتين
hāʾulāʾi-l-ʾimraʾatain

c. Those two men
أولئك الرّجلين
ʾūlāʾika-r-rağulain

d. Those two women
أولئك الإمرأتين
ʾūlāʾika-l-ʾimraʾatain

e. The two men who smile
الرّجلان الذي يَبتسمانِ
ʾar-rağulāni-l-laḏī yabtasimān

f. The two women who smile
الإمرأتان الذي تَبتسمانِ
ʾal-ʾimraʾatāni-l-laḏī tabtasimān

g. They are two good boys.
هم ولدان جيدون.
hum waladāni ğayyidūn

h. They are two good girls.
هم بنتان جيداتُ.
hum bintāni jayyidāt

i. Two active, diligent women attended the meeting.
إمرأتان مجتهدات نشيطات، حَضرتَا الإجتماع.
ʾimraʾatāni muğtahidāt našīṭāt ḥaḍaratā-l-ʾiğtimāʿ

j. Two active, diligent men attended the meeting.
رجلان مجتهدون نشيطون، حَضرا الإجتماع.
rağulān muğtahidūni našīṭūna ḥaḍarā-l-ʾiğtimāʿ

k. The two girls love their mother.
البنتان تَحْبّانِ أمّهم.
ʾal-bintāni tuḥibbāni ʾummahum

l. Both boys love football.

كلا الأولاد يَحْبّونَ كرة قدمَ.
kilā-l-ʾawlādi yuḥibbūna kurata qadam

m. Both girls love football.
كلتا البنات يَحْببنَ كرة قدمَ.
kiltā-l-banāti yuḥbibna kurata qadam

n. Both of them love football.
كلاهما يَحْبُّ كرة قدم.
kilāhumā yuḥibbu kurata qadam

o. The boy and the girl are happy.
إنّ الولدَ والبنتَ سعيدةٌ.
ʾinnna-l-walada wa-l-binta saʿidah

p. She and I were school fellows.
هي وأنا كُنْتُ زملاء دراسةٍ.
hiya wa ʾanā kuntu zumalāʾa dirāsah

q. He and his wife are always fighting.
هو وزوجته تُقاتلُ دائما.
huwa wa zawǧatuhu tuqātilu dāʾimā

r. Jack and Jane went to the garden and played football.
جاك وجان ذهبَ إلى الحديقةِ ولعبَ كرة قدم.
ǧāk wa ǧān ḏahaba ʾilā-l-ḥadīqati wa laʿiba kurata qadam

s. Jack and John went to the garden and played football.
جاك وجون ذهبا إلى الحديقةِ ولعبا كرة قدم.
ǧāk wa ǧūn ḏahabā ʾilā-l-ḥadīqati wa laʿibā kurata qadam

t. He finished the work after a couple of hours.
انهىَ العملَ بعد ساعتين.
ʾanhā-l-ʿamala baʿda sāʿatain

u. A couple of boys were playing in the garden.
زوج الأولادِ كَانَ يَلْعبُ في الحديقةِ.
zawǧu-l-ʾawlādi kāna yalʿabu fī-l-ḥadīqah

v. A couple of girls were playing in the garden.
زوج البناتِ كَانَ يَلْعبُ في الحديقةِ.
zawǧu-l-banāti kāna yalʿabu fī-l-ḥadīqah

w. A couple of birds were flying over there.
زوج الطّيورِ كَانَ يَطِيرُ هناك.
zawǧu-ṭ-ṭuyūri kāna yaṭīru hunāk

x. He drank a couple of glasses.
شَربَ زوجَ الأقداح.
šariba zawǧa-l-ʾaqdāḥ

Test Analysis:

The system recognizes the dual only when the noun is preceded by *two* and, only in this case, succeeds in assigning the correct number to the verbs as in

examples *e, f, i, j* and *k*. In examples *a* through *d* the system fails in getting the dual form of the demonstratives. In examples *e* and *f* the system fails in getting the dual form of the relative pronouns. In examples *g* through *j* the system fails in getting the dual form of adjectives. In example *k* the anaphoric pronominal *their* was not interpreted as dual. The noun modified by *both* must be dual and the verb must be singular. However, in examples *l* and *m* this rule is not correctly followed. In example *n* the translation is correct. When two singular nouns are joined by the coordinating conjunction *and* they must be interpreted as dual. This rule is not followed in examples *o* through *r*. In example s it is not clear why the system has succeeded in identifying the dual. Perhaps this is because it can identify the gender of both proper nouns (*Jack* and *John*). This may explain the contrast between examples *r* and *s*. In example *t* the system identifies *couple of* as meaning dual, yet in examples *u* through *x*, it fails in reaching this conclusion.

6.4.1. Proposed Solution

The problem with the dual can be dealt with during the analysis phase. Instead of marking the NP as *plural* the system must make allowance for *dual*, taking into consideration that the target language is Arabic. Then during generation the system will specify the grammatical changes needed for agreement and the final morphological realization.

If the NP is marked as *plural* with no allowance for *dual*, then the system will have, during transfer, to look back again at the internal structure of the NP and whether it denotes two entities or more. This means that the transfer phase will make further analysis of the source language, which is not related to its original task, that is making bilingual changes. Therefore despite the fact that marking the NP for dual in analysis is not linguistically motivated, no other phase has the ability or efficiency to undertake this task. The monolingual analyzer has the ability to look forward and backward (backtracking) and search for contextual clues to define the correct number of the noun, an advantage which is lacking in other phases.

6.5. Anaphora

Anaphora is a "reference being made to an entity mentioned explicitly elsewhere in a text."[1] It covers the area of reflexive pronouns, reciprocals, possessives, and pronominals. "When translating into languages which mark the gender of pronouns, for example, it is essential to resolve the anaphoric relations,"[2] otherwise, agreement rules (between anaphoric pronoun and antecedent) in the target language will be violated resulting in an incoherent, ungrammatical, and sometime incomprehensible translation.

It must be noted that "not all pronouns are anaphoric"[3]. The pronouns *I*, *we* and *you* are not anaphoric as they "refer directly to speaker/writer and addressee(s) rather than by virtue of anaphoric relation to an antecedent."[4] This is why I assigned a separate section above for discussing the problems which pronouns may pose to MT. Pronouns were discussed earlier as a translational problem, i.e. getting the correct equivalent for each pronoun. In this section I will discuss some pronouns as anaphoric pronominals, i.e. establishing the link between the pronoun and its antecedent.

Test Examples: Reflexive Anaphors: *Himself, Herself* and *Ourselves*.

a. The boy must depend on himself.
الولد يَجِبُ أنْ يَعتمدَ على نفسه.
ʾal-walad yağibu ʾan yaʿtamida ʿalā nafsih

b. The girl must depend on herself.
البنت يَجِبُ أنْ تَعتمدَ على نفسها.
ʾal-bint yağibu ʾan taʿtamida ʿalā nafsihā

c. We must depend on ourselves.
نحن يَجِبُ أنْ نَعتمدَ على أنفسنا.
naḥnu yağibu ʾan naʿtamida ʿalā ʾanfusinā

Test Analysis:

The system makes correct translations of the reflexives. It is expected that there will be no problem in translating the reflexive pronouns *himself* and *herself* as the English pronouns convey both gender and number. Similarly, the

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 95.
[2] Ibid., p. 95.
[3] Rodney Huddleston, *An Introduction to English Transformational Syntax*, Essex: Longman, 1976, p. 252.
[4] Ibid.

reflexive *ourselves* needs no further analysis because the Arabic translation is straightforward.

The problem arises when translating *yourself, yourselves, themselves* and *itself*. In Arabic, the anaphoric pronoun which is equivalent to *yourself* must agree with the antecedent in gender (feminine or masculine):

yourself (m)	=>	nafsuka[1]
yourself (f)	=>	nafsuki[2]

Test Examples: Reflexive Anaphors: *Yourself*

a. The man said to his son, "You must depend on yourself."
قالَ الرّجلُ أنْ إبنه، " أنتَ يَجِبُ أنْ تَعتمدَ على نفسك. "
qāla-r-rağulu ʾan ʾibnahu, "ʾanta yağibu ʾan taʿtamida ʿalā nafsik."

b. The man said to his daughter, "You must depend on yourself."
قالَ الرّجلُ أنْ بنته، " أنتَ يَجِبُ أنْ تَعتمدَ على نفسك. "
qāla-r-rağulu ʾan bintahu, "ʾanta yağibu ʾan taʿtamida ʿalā nafsik."

Test Analysis:

Only the translation in example *a* is correct. In example *b* the system fails in identifying the correct gender of the pronoun *you* and the reflexive *yourself*.

When translating the reflexive *yourselves* we need first to detect its appropriate gender (masculine or feminine) and number (dual or plural):

yourselves (dual)	=>	ʾanfusukumā[3]
yourselves (plural, masc)	=>	ʾanfusukum[4]
yourselves (plural, fem)	=>	ʾanfusukunna[5]

Test Examples: Reflexive Anaphors: *Yourselves*

a. The man said to his two sons, "You must depend on yourselves."
قالَ الرّجلُ أنْ أبنائه، " أنتَ يَجِبُ أنْ تَعتمدَ على أنفسكم. "
qāla-r-rağulu ʾan ʾabnaʾihi, "ʾanta yağibu ʾan taʿtamida ʿalā anfusikum."

b. The man said to his sons, "You must depend on yourselves."

[1] Word ending may differ according to case. For example, *nafsuka* may also be realized as *nafsaka* (ACC) and *nafsika* (GEN).
[2] nafsaki (ACC), nafsiki (GEN).
[3] ʾanfusakumā (ACC), ʾanfusikumā (GEN)
[4] ʾanfusakum (ACC), ʾanfusikum (GEN).
[5] ʾanfusakunna (ACC), ʾanfusikunna (GEN).

قالَ الرّجلُ أنْ أبنائه، " أنتَ يَجِبُ أنْ تَعتمدَ على أنفسكم. "
qāla-r-rağulu ʾan ʾabnaʾihi, "ʾanta yağibu ʾan taʿtamida ʿalā anfusikum."

c. The man said to his daughters, "You must depend on yourselves."
قالَ الرّجلُ أنْ بناته، " أنتَ يَجِبُ أنْ تَعتمدَ على أنفسكم. "
qāla-r-rağulu ʾan ʾbanātihi, "ʾanta yağibu ʾan taʿtamida ʿalā anfusikum."

Test Analysis:

The translation of the pronoun *you* is wrong in the three examples above, resulting in clear violation of the agreement rules between reflexives and their antecedents. In example *a*, it is not clear why the system fails in identifying *two sons* as dual.

When translating the reflexive *themselves* we need first to detect its appropriate gender (masculine or feminine) and number (dual or plural) as well as whether it denotes human or non-human entity. When the antecedent in non-human plural it is treated as feminine and singular.

themselves (dual)	=>	nafsahumā[1]
themselves (plural, masc)	=>	ʾanfusuhum[2]
themselves (plural, fem)	=>	ʾanfusuhunna[3]
themselves (plural, non-human)	=>	nafsuhā[4]

Test Examples: Reflexive Anaphors: *Themselves*

a. The two boys must depend on themselves.
الولدان يَجِبُ أنْ يَعتمدا على أنفسهم.
ʾal-waladāni yağibu ʾan yaʿtamidā ʿalā ʾanfusihim.

b. The two girls must depend on themselves.
البنتان يَجِبُ أنْ تَعتمدا على أنفسهم.
ʾal-bintāni yağibu ʾan taʿtamidā ʿalā ʾanfusihim.

c. The boys must depend on themselves.
الأولاد يَجِبُ أنْ يَعتمدوا على أنفسهم.
ʾal-ʾawlādu yağibu ʾan yaʿtamidū ʿalā ʾanfusihim.

d. The girls must depend on themselves.
البنات يَجِبُ أنْ يَعتمدنَ على أنفسهم.
ʾal-banātu yağibu ʾan yaʿtamidna ʿalā ʾanfusihim.

[1] nafsaihimā (ACC, GEN).
[2] ʾanfusahum (ACC), ʾanfusihim (GEN).
[3] ʾanfusahunna (ACC), ʾanfusihinna (GEN).
[4] nafsahā (ACC), nafsihā (GEN)

e. The cats must depend on themselves.
القطط يَجِبُ أنْ تَعتمدَ على أنفسهم.
ʾal-qiṭaṭu yağibu ʾan taʿtamida ʿalā ʾanfusihim.

Test Analysis:
Only translation in example *c* is correct. In all other examples the reflexives do not agree in number and gender with their antecedents.

Regarding the anaphoric reflexive pronoun *itself*, the Arabic equivalent must agree with the antecedent in gender.

itself (fem)	=>	nafsuhā
itself (masc)	=>	nafsuhu[1]

Test Examples: Reflexive Anaphors: *Itself*

a. The cow must depend on itself.
البقرة يَجِبُ أنْ تَعتمدَ على نفسها.
ʾal-baqaratu yağibu ʾan taʿtamida ʿalā nafsihā.

b. The bull must depend on itself.
الثّور يَجِبُ أنْ يَعتمدَ على نفسه.
ʾaṯ-ṯawru yağibu ʾan yaʿtamida ʿalā nafsihi.

Test Analysis:
The system successfully identifies the correct gender of *itself* in both examples.

Regarding reciprocals, Arabic makes a distinction according to number and gender as follows.

each other (dual, masc)	=>	kullun minhumā-l-ʾāḫar
each other (dual, fem)	=>	kullun minhumā-l-ʾuḫrā
each other (plural, masc)	=>	kullun minhumu-l-ʾāḫar
each other (plural, fem)	=>	kullun minhunna-l-ʾuḫrā

Test Examples: Reciprocal Anaphors: *Each Other*

a. The two boys love each other.
الولدان يَحْبّانِ كل منهم الآخر.
ʾal-waladāni yuḥibbāni kullun minhumu-l-ʾāḫar.

b. The two girls love each other.
البنتان تَحْبّانِ كل منهم الآخر.
ʾal-bintāni tuḥibbāni kullun minhumu-l-ʾāḫar.

c. The boys love each other.

[1] nafsahu (ACC), nafsihi (GEN)

الأولاد يَحْبّونَ كل منهم الآخرَ.
ʾal-ʾawlādu yuḥibbūna kullun minhumu-l-ʾāḫar.

d. The girls love each other.
البنات يَحْببنَ كل منهم الآخرَ.
ʾal-banātu yuḥbibna kullun minhumu-l-ʾāḫar.

Test Analysis:

Only the translation in *c* is correct. In all other examples the system fails in assigning the correct number and/or gender for the reciprocal to make it agree with its antecedent.

Another type of anaphora is the pronominals, which can have either internal or external reference, i.e. their reference may or may not be mentioned in the sentence. In the following example, the most workable explanation is that pronouns have their reference in the sentence.

Test Examples: Anaphora: Pronominals

a. The boys believe that they are diligent.
إنّ الأولادَ يَعتقدونَ بأنّهم مجتهدون.
ʾinna-l-ʾawlāda yaʿtaqidūna bi-ʾannahum muǧtahidūn.

b. The two boys believe that they are diligent.
إنّ الولدين يَعتقدانِ بأنّهم مجتهدون.
ʾinna-l-waladaini yaʿtaqidāni bi-ʾannahum muǧtahidūn.

c. The two girls believe that they are diligent.
إنّ البنتين تَعتقدانِ بأنّهم مجتهدون.
ʾinna-l-bintain taʿtaqidāni bi-ʾannahum muǧtahidūn.

d. The girls believe that they are diligent.
إنّ البناتِ يَعتقدنَ بأنّهم مجتهدون.
ʾinna-l-banāti yaʿtaqidna bi-ʾannahum muǧtahidūn.

e. The girls went to bed because they were tired.
البنات نِمنَ لأن هم قَدْ أتعبوا.
ʾal-banātu nimna liʾanna hum qad ʾutʿibū.

f. The girls met their brother while they were walking in the garden.
البنات قَابلنَ أخَاهم بينما هم كَانوا يَمْشونَ في الحديقةِ.
ʾal-banātu qabalna ʾaḫāhum bainamā hum kanū yamšūna fī-l-ḥadīqah.

g. The girls met their brothers while they were walking in the garden.
البنات قَابلنَ أخوتهم بينما هم كَانوا يَمْشونَ في الحديقةِ.
ʾal-banātu qabalna ʾiḫwatahum bainamā hum kanū yamšūna fī-l-ḥadīqah.

h. The boy depends on his father.
يَعتمدُ الولدُ على أبيه.
yaʿtamidu-l-waladu ʿalā ʾabīh.

i. The girl depends on her father.

تَعتمدُ البنتُ على أبيها.
taʿtamidu-l-bintu ʿalā ʾabīhā.

j. The boys depend on their father.
الأولاد يَعتمدونَ على أبيهم.
ʾal-ʾawalādu yaʿtamidūna ʿalā ʾabīhim.

k. The two boys depend on their father.
الولدان يَعتمدانِ على أبيهم.
ʾal-waladāni yaʿtamidāni ʿalā ʾabīhim.

l. The two girls depend on their father.
البنتان تَعتمدانِ على أبيهم.
ʾal-bintāni taʿtamidāni ʿalā ʾabīhim.

m. The girls depend on their father.
البنات يَعتمدنَ على أبيهم.
ʾal-banātu yaʿtamidna ʿalā ʾabīhim.

n. The girls and their brother are diligent.
إنّ البناتِ وأخَاهم مجتهدُ.
ʿinna-l-banāti wa ʾaḫāhum muǧtahid.

o. The cats depend on their father.
القطط تَعتمدُ على أبيهم.
ʾal-qiṭaṭ taʿtamidu ʿalā ʾabīhim.

p. The bull depends on its father.
يَعتمدُ الثّورُ على أبيه.
yaʿtamidu-ṯ-ṯawru ʿalā ʾabīh.

q. The cow depends on its father.
تَعتمدُ البقرةُ على أبيها.
taʿtamidu-l-baqaratu ʿalā ʾabīhā.

r. The book and its cover were torn.
الكتاب وغطائه قَدْ مُزّقا.
ʾal-kitābu wa ġiṭāʾihi qad muzziqā.

s. The table and its cover were cleaned.
المنضدة وغطائها قَدْ نُظّفا.
ʾal-minḍadatu wa ġiṭāʾihā qad nuẓẓifā.

Test Analysis:

The anaphoric pronominal *they* is translated correctly in example *a*; yet in examples *b* through *f*, the pronouns clearly do not agree with their antecedents in number and/or gender. In example *g* the pronoun is ambiguous as it can refer either to *the girls* or *their brothers*. The same problem is shown with the anaphoric possessive *their*. Only the translation in *j* is correct; yet in examples *k* through *o* the pronouns in Arabic do not agree with their antecedent in number and/or gender. Examples *p* through *s* show how the system

successfully assigns the gender of the pronoun. It is noted that in examples *r* and *s* the coordinated words *ġiṭāʾihi* and *ġiṭāʾihā* are not assigned correct case.

Pronominals may also have external reference, i.e. their antecedent is not mentioned in the sentence. The antecedent may have been mentioned in earlier sentences, or it may just be "'on the mind' of both the speaker and the hearer."[1]

Test Examples: Pronominals with External References

a. They depend on the boy.
يَعتمدونَ على الولدِ.
yaʿtamidūna ʿalā-l-waladi.

b. The boy depends on their father.
يَعتمدُ الولدُ على أبيهم.
yaʿtamidu-l-waladu ʿalā ʾabīhim.

c. The boy depends on them.
يَعتمدُ الولدُ عليهم.
yaʿtamidu-l-waladu ʿalaihim.

Test Analysis:

The anaphoric pronouns in the above examples have external references as there are no antecedents in the sentences. So the best solution, which the system has already made, is to use the default unmarked form, i.e. the masculine.

6.5.1. Proposed Solutions

1. With the reflexive pronouns *himself*, *herself*, and *ourselves*, no special handling or further analysis is needed since they have straightforward Arabic equivalents. They can be taken as a free ride.
2. The other reflexive pronouns *yourself*, *yourselves, themselves, itself* as well as pronominals must be co-indexed with their antecedents and assigned appropriate grammatical features of number and gender.
3. Sometimes the anaphoric reference cannot be established due to ambiguity. "In fact anaphora can be thought of as a sort of ambiguity, in that the antecedent of a given pronoun might be uncertain."[2] In this case it is, to a great extent, considered the fault of the writer who

[1] Ian Roberts, *Comparative Syntax*, London: Arnold, 1997, p. 128.
[2] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 95.

"should eliminate such ambiguity, either by substituting a noun for the pronoun or by clarifying the antecedent."[1]

Test Example: Ambiguous Antecedent

a. The men met the women and they were happy.
الرّجال قَابلوا النِّساءَ وهم كَانوا سعيداتَ.
ʾar-riğālu qābalū-n-nisāʾa wa hum kānū saʿīdāt.

In the above example, the pronoun *they* can refer to *men* or *women* or both of them or even an external referent. Neither structural, textual, nor semantic clues can be found to resolve this ambiguity. Only the real world context (which is not accessible to the machine) can provide an answer. Yet the machine can take a best-guess approach to give a plausible output.

6.6. Infinitival Phrases

Infinitival phrases are composed of verbs in the infinitive form usually preceded by the infinitival particle *to*. They lack overt subject, tense and agreement, or as Radford puts it: "Nonfinite verb-forms are intrinsically tenseless and agreementless."[2] Deep structure analysis of these phrases, however, is "needed for translation into a language which makes the subject of the embedded sentence explicit or where knowing what the subject would be is needed for gender or number agreement."[3]

Infinitival phrases are usually classified into three categories: raising sentences, control sentences, and plain infinitives (infinitives without the particle *to*).

6.6.1. Raising Sentences

Raising sentence involves sentences where the subject has moved from the infinitive complement to become the subject of the main sentence. We can first look at the following examples:

It seems that he is a good student.
He_i seems e_i to be a good student.

[1] Perrin Smith Corder, *Handbook of Current English*, 3rd ed., Glenview, Illinois: Scott, Foresman and Company, 1968, p. 90.
[2] Andrew Radford, *Transformational Grammar: A First Course*, Cambridge: Cambridge University Press, 1988, p. 288.
[3] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 32.

The second sentence has undergone ‘raising’ where the subject of the complement clause is “raised up to become the matrix clause subject by application of **(subject to subject) raising**, leaving behind an empty category trace.”[1] In the above sentence we notice that the empty category *e* is co-indexed with the subject *he*. This process occurs with a limited category of verbs (called ‘raising verbs’) like *seem, happen, appear* and *turn out*. It occurs also with “a small class of adjectives, which we can call raising adjectives”[2] such as *likely* and *certain*.

With raising sentences, the subject of the infinitive is the subject of the main clause. So in the Arabic translation the verb in the second clause must agree in number and gender with the subject in the first clause.

Test Examples: Raising Sentences

a. The boy seems to be happy.
يَظْهرُ الولدُ أنْ يَكُونَ سعيدَ.
yaẓharu-l-waladu ʾan yakūna saʿīd.

b. The girl seems to be happy.
تَظْهرُ البنتُ أنْ تَكُونَ سعيدةَ.
taẓharu-l-bintu ʾan takūna saʿīdah.

c. The boys turned out to be smart.
الأولاد ظهروا أذكياء.
ʾal-ʾawlādu ẓaharū ʾaḏkiyāʾ.

d. The girls turned out to be smart.
البنات ظهرنَ ذكياتَ.
ʾal-banātu ẓaharna ḏakiyyāt.

e. The boy is believed to be good.
إنّ الولدَ يَعتقدُ أنْ يَكُونَ جيدَ.
ʾinna-l-walada yaʿtaqidu ʾan yakūna ǧayyid.

f. The girl is believed to be good.
إنّ البنتَ تَعتقدُ أنْ تَكُونَ جيدةَ.
ʾinna-l-binta taʿtaqidu ʾan takūna ǧayyidah.

g. The boys appear to be good.
الأولاد يَبْدونَ الجيدين.
ʾal-ʾawlādu yabdūna-l-ǧayyidīn.

h. The girls appear to be good.

---

[1] Andrew Radford, *Syntax: A minimalist introduction*, Cambridge: Cambridge University Press, 1997, pp. 175-6.
[2] Robert Borsley, *Syntactic Theory: A Unified Approach*, 2nd ed., London: Arnold, 1999, p. 157.

البنات يَبْدونَ الجيداتِ.
ʾal-banātu yabdūna-l-ǧayyidāt.

i. The boys happened to be smart.
صَادفوا الأولادُ وأن كان أذكياء.
ṣādafū-l-ʾawlādu wa ʾan kāna ʾaḏkiyāʾ.

j. The girls happened to be smart.
صَادفنَ البناتُ وأن كان ذكياتُ.
ṣādafna-l-banātu wa ʾan kāna dakiyyāt.

k. The girl is likely to be happy.
إنّ البنتَ محتمل أن يكون سعيدة.
ʾinna-l-binta muḥtamalun ʾan yakūna saʿīdah.

l. The girls are certain to be happy.
إنّ البناتِ بعض أنْ يَكُنّ سعيداتَ.
ʾinna-l-banāti baʿḍ ʾan yakunna saʿīdāt.

Test Analysis:

In examples *a* through *h* the system successfully identifies the subject of the raising verb to be the subject of the infinitive. However, in the rest of the examples the system fails in detecting the correct subject. It is not clear why the word order in the translation of the verb *seem* (in examples *a* and *b*) is different from that in the translation of the verb *turn out* (in examples *c* and *d*). It is not clear either why the adjective in *g* is prefixed by the definite article. Perhaps the agreement is done wrongly with the noun instead of the pronoun.

6.6.2. Control Sentences

Although control sentences have superficially similar structures with raising sentences in that they both take infinitival complements, they have different underlying structure.

He_i seems e_i to understand the question.
He_i tries to PRO_i understand the question.

While the subject of the infinitival phrase in raising sentences is a trace of movement to the subject position of the raising verb, the subject of the infinitival phrase in control sentences is a PRO, which is "a referential pronoun which takes its reference from its controller."[1] The controller can be the subject of the control verb as in the above example, or it can be its object as in the following example:

He wants $John_i$ to PRO_i understand the question.

[1] Andrew Radford, *Syntax: A minimalist introduction*, Cambridge: Cambridge University Press, 1997, p. 179.

The controller can also be the object of preposition as in the following example:

He appealed to $John_i$ to PRO_i understand the question.

Moreover, the controller may be vague and cannot be identified. This occurs when "there is no overt controller within the same sentence and the reference of the controllee is therefore arbitrary."[1]

PRO_i To quit the job now will be wrong.

Here the reference "has an 'arbitrary' interpretation meaning approximately 'someone or other' … with a weak implication that the speaker and hearer are included."[2]

The controller may also be implied and not explicitly mentioned in the sentence as in the following example, where the reference is the inferred pronoun *I*:

My aim was PRO_i to make money.

Trying to account for the different behavior of both raising and control sentences, Borsley[3] proposes that the basic difference between the two categories is that control verbs, like *try*, are capable of assigning thematic roles (or θ-roles for short) like agent and patient while raising verbs are not able to θ-mark their subject. This is evident in the fact that raising verbs, and not control verbs, can take the dummy expletive pronoun *it* in the subject position:

It seems that John understands the question.
* It tries that John understands the question.

This is why the subject of the infinitival phrase in raising sentences is interpreted as the trace to the subject of the raising verb. Indeed, it is the verb in the infinitival phrase that assigns the θ-role AGENT to the subject. This interpretation, however, is not working with control sentences. A control verb θ-marks its subject and hence cannot receive a θ-role from the verb of the infinitival phrase because this will violate the θ-criterion which states that: "Each argument bears one and only one θ-

[1] Geoffrey Horrocks, *Generative Grammar*, London: Longman, 1987, p. 254.
[2] Ian Roberts, *Comparative Syntax*, London: Arnold, 1997, p. 135
[3] Robert Borsley, *Syntactic Theory: A Unified Approach*, 2nd ed., London: Arnold, 1999, p. 161.

role, and each θ-role is assigned to one and only one argument."[1] The only feasible interpretation of the subject is then a PRO which refers back to its controller.

Test Examples: Control Sentences

a. He persuaded her to visit his father.
اقنعها أنْ تَزُورَ أبوه.
ʾaqnaʿahā ʾan tazūra ʾabūh.

b. He wanted to visit his father.
ارادَ أنْ يَزُورَ أبوه.
ʾarāda ʾan yazūra ʾabūh.

c. He wanted her to visit his father.
ارادها أنْ تَزُورَ أبوه.
ʾarādahā ʾan tazūra ʾabūh.

d. He appealed to her to visit his father.
نَاشدَ إلىها أنْ تَزُورَ أبوه.
nāšada ʾilā hā ʾan tazūra ʾabūh.

e. She went to England to learn English.
ذَهبتْ إلى إنجلترا أنْ تَتعلّمَ إنكليزية.
ḏahabat ʾilā ʾingiltirā ʾan tataʿallam inkilīziyyah.

f. She is able to go to the party.
هي قادرةٌ أنْ تَذْهبَ إلى الحفلة.
hiya qādirah ʾan taḏhaba ʾilā-l-ḥaflah.

g. He promised her to visit his father.
وَعدها أنْ تَزُورَ أبوه.
waʿadahā ʾan tazūra ʾabūh.

h. My aim was to make money.
هدفي كَانَ أنْ يَجْمعَ مالاً.
hadafī kāna ʾan yağmaʿa mālā.

i. His aim was to make money.
هدفه كَانَ أنْ يَجْمعَ مالاً.
hadafuhu kāna ʾan yağmaʿa mālā.

j. Her aim was to make money.
هدفها كَانَ أنْ يَجْمعَ مالاً.
hadafuhā kāna ʾan yağmaʿa mālā.

k. The task of the manager is to organize work.
إنّ مهمة المديرِ أنْ تُنظّمَ عمل.
ʾinna muhimmata-l-mudīri ʾan tunaẓẓima ʿamal.

l. He thought that his responsibility was to bring peace to the region.

[1] N. Chomsky, *Lectures on Government and Binding*, Dordrecht: Foris, 1981, p. 36.

اعتقدَ بأنّ مسؤوليته كانتْ أنْ تَجْلبَ سلام إلى المنطقةِ.
ʾiʿtaqada biʾanna masʾūliyyatahu kānat ʾan tağliba salām ʾilā-l-manṭiqah.

m. We are not in a hurry to know the results.
نحن لا في عجلةِ أنْ تَعْرفَ النَّتائِج.
naḥnu lā fī ʿağalah ʾan taʿrifa-n-natāʾiğ.

n. They have nothing to do.
هم قَدْ لاشيئَ أنْ يَعمَلُ.
hum qad lā šaiʾu ʾan yaʿmal.

o. She found nothing to eat.
وَجدتْ لاشيئَ أنْ يَأكلَ.
wağadat lā šaiʾa ʾan yaʾkul.

p. It was silly of them to come late.
كان سخيف منهم أنْ يَجيءَ متأخّر.
kāna saḫīf minhum ʾan yağīʾa mutaʾaḫḫir.

q. To quit the job now will be wrong.
أنْ تَتْركَ الشّغل الآن سَيَكُونُ خاطئَ.
ʾan tatruka-š-šuġla-l-ʾāna sayakūnu ḫāṭiʾ.

r. There was nothing to eat in the fridge.
هناك قَدْ لاشيئَ أنْ يَأكلَ في الثلاجةِ.
hunāka qad lā šaiʾa ʾan yaʾkula fī-ṯ-ṯallāğah.

Test Analysis:

In examples *a* through *f* the system succeeds in detecting the correct subject of the infinitival. However, in examples *g* through *p*, it fails in identifying the correct subject, and so fails in making the verb agree with its appropriate subject. In examples *q* and *r* the subject could not be inferred from the sentence. Maybe the best solution is to translate them as verbal nouns (maṣdars) to preserve the ambiguity of the subject.

6.6.3. Plain (without *to*) Infinitives

Plain infinitives, or infinitives without the particle *to*, occur after modal verbs. In this case the modal verb is an auxiliary verb while the main verb is the infinitive.

Test Examples: Plain Infinitives

a. He can speak English.
هو يُمكنُ أنْ يَتكلّمَ إنجليزي.
huwa yumkinu ʾan yatakallami ʾinğilīzī.

b. She can speak English.
هي يُمكنُ أنْ تَتكلّمَ إنكليزية.

hiya yumkinu ʾan tatakallama ʾinkilīziyyah.

c. The boys could understand him.
الأولاد يُمكنُ أنْ يَفْهموه.
ʾal-ʾawlādu yumkinu ʾan yafhamūh.

d. The girls could understand him.
البنات يُمكنُ أنْ يَفْهمنه.
ʾal-banātu yumkinu ʾan yafhamnah.

e. The boys should help him.
الأولاد يَجِبُ أنْ يُساعدوه.
ʾal-ʾawlādu yağibu ʾan yusāʿidūh.

f. The girls should help him.
البنات يَجِبُ أنْ يُساعدنه.
ʾal-banātu yağibu ʾan yusāʿidnah.

g. I will help you.
أنا سَأُساعدكَ.
ʾanā saʾusāʿiduk.

h. You must go now.
أنتَ يَجِبُ أنْ تَذْهبَ الآن.
ʾanta yağibu ʾan taḏhaba-l-ʾān.

Test Analysis:

In all examples the system succeeds in making the translation of plain infinitives agree with their subjects. It is noted that the translations of the word *English* in the first two examples agree with the subject in gender, which is not correct, as the word is used as a noun not an adjective. It is also noted that the word is spelled differently each time (ʾinjilīzī vs. ʾinkilīziyyah).

The plain infinitives can also occur after verbs of perception and after *make* and *let*. In this case the infinitives are known to originally "come from embedded sentences"[1] and their subjects are always the objects of the main sentences.

Test Examples: Plain Infinitives

a. I saw the boy carry the book.
رَأيتُ الولدَ يَحْملُ الكتابَ.
raʾaitu-l-walada yaḥmilu-l-kitāb.

b. I saw the girl carry the book.
رَأيتُ البنتَ تَحْملُ الكتابَ.
raʾaitu-l-binta taḥmilu-l-kitāb.

[1] Mark Lester, *Introductory Transformational Grammar of English*, 2nd ed., New York: Holt, Rinehart and Winston, 1976, p. 240.

c. They heard her speak.
سَمعوها تَتكلّمُ.
samiʿūhā tatakallam.

d. They heard him speak.
سَمعوه يَتكلّمُ.
samiʿūhu yatakallam.

e. Don't make me laugh.
لا تَجْعلني اَضْحكُ.
lā tağʿalnī aḍḥak.

f. He let the boys use his car.
تَركَ الأولادَ يَستعملونَ سيارته.
taraka-l-ʾawlāda yastaʿmilūna sayyāratah.

g. He let the girls use his car.
تَركَ البناتِ يَستعملنَ سيارته.
taraka-l-banāti yastaʿmilna sayyāratah.

Test Analysis:

The system also succeeds here in making the translation of plain infinitives agree with their subjects.

It must be noted that there is a non-finite verb construction other than the infinitive. This is the gerund (verb + ing) which follows some verbs such as *avoid, consider, delay, deny, enjoy, forbid, mind, postpone, risk* and *suggest*. The gerund following these verbs functions as a noun and is "used as the object"[1] of the verb. I did not allocate a separate section for this type because the translation into Arabic does not need linking these verb forms with controllers. The best solution is to translate them into similar non-finite forms, which is called *maṣdar* or verbal noun.

Test Examples: The Gerund

a. I can't avoid meeting him.
أنا لا يُمكنُ أنْ أتجنّبَ أُقابله.
ʾanā lā yumkinu ʾan ʾatağannaba ʾuqābilah.

b. I will consider visiting him.
أنا سَأعتبرُ زيَارَته.
ʾanā saʾaʿtabiru ziyāratah.

c. He denied attacking her.
انكرَ مُهَاجَمَتَهَا.
ʾankara muhāğamata hā.

[1] Betty Schrampfer Azar, *Understanding and Using English Grammar*, Englewood, New Jersey: Prentice-Hall, 1989, p. 150.

d. I can't risk losing that job.
أنا لا يُمكنُ أنْ أُخاطرَ فَقد ذلك الشّغل.
ʾanā lā yumkinu ʾan ʾuḫāṭira faqda ḏalika-š-šuġl.

Test Analysis:

In the first example the system translates the gerund as a verb while in the rest of the examples, it translates them as *maṣdar*s. Yet all the sentences are considered ungrammatical because they lack the particle *ʾann* in example *a*, lack the preposition *bi* in example *d*, have wrong morphological generation in example *c*, or make a wrong translation of the verb in example *b*.

6.6.4. Proposed Solution

Making correct agreement in the target language depends on detecting the right subject of the verb in the infinitival phrase. This can be done by properly co-indexing the trace (in raising sentences) or PRO (in control sentences) with the appropriate NP.

6.7. Participial Phrases

Like infinitival phrases, participial phrases are both subjectless and tenseless. Participial phrases can be composed of a present participle:

Wishing to earn more money, he traveled abroad.

or a past participle:

Hardened by the experience, he did not give up.

or compound (present followed by past) participles:

Having completed his work, the old man went home.

Participial phrases pose a problem during translation especially when the target language, like Arabic, has no equivalent structure and hence requires an explicit subject and subject-verb agreement.

Test Examples: Participial Phrases

a. Wishing to earn more money, he traveled abroad.
الرَّغْبَة أنْ تَكْسبَ فلوس أكثر، سَافرَ في الخارج.
ʾar-raġbatu ʾan taksaba fulūs ʾakṯar, sāfara fī-l-ḫāriğ.

b. Going to school, the student lost his bag.
تَذْهبُ إلى المدرسةِ، الطّالب فَقدَ حقيبته.
taḏhabu ʾilā-l-madrasah, aṭ-ṭalibu faqada ḥaqībatah.

c. Born in England, she spoke perfect English.
المولد في إنجلترا، تَكلّمتْ إنكليزية تامة.

ʾal-mawlid fī ʾingiltirā, takallamat ʾinklīziyyah tāmmah.

d. Hardened by the experience, he did not give up.
مُصلَّب بالخبرةِ، هو مَا استسلمَ.
muṣallab bi-l-ẖibrah, huwa mā ʾistaslama.

e. Taken by surprise, she gave up.
أخذ على حين غرة، استسلمتْ.
ʾaẖaḏa ʾalā ḥīni ġirrah, ʾistaslamat.

f. Having completed his work, the old man went home.
بعد ما اكملَ عمله، ذهبَ الرجل العجوزُ الى البيت.
baʿda mā ʾakmala ʿamalah, ḏahaba-r-rağulu-l-ʿağūzu ʾilā-l-bait.

g. Having completed her work, the old woman went home.
بعد ما اكملَ عملها، ذهبَ العجوزةُ الى البيت.
baʿda mā ʾakmala ʿamalahā, ḏahaba-l-ʿağuzatu ʾilā-l-bait.

h. Having been abroad for thirty years, they knew many languages.
بعد ما كانَ في الخارج لثلاثون سنةِ، عَرفوا العديد من اللّغاتِ.
baʿda mā kāna fī-l-ẖariğ li-ṯalāṯūna sanati, ʿarafū-l-ʿadīda mina-l-luġāti.

i. Having been abroad for thirty years, she knew many languages.
بعد ما كانَ في الخارج لثلاثون سنةِ، عَرفتْ العديد من اللّغاتِ.
baʿdamā kāna fī-l-ẖariğ li-ṯalāṯūna sanati, ʿarafati-l-ʿadīda mina-l-luġāti.

Test Analysis:

Only translation in example *f* is correct. In all other examples the translation is wrong either because the verb is translated as a nominal or as a verb that lacks correct agreement features.

6.7.1. Proposed Solution

In order to render this type of sentence into Arabic, the sentence needs to be rephrased. In doing so, it must be taken into consideration that "the performer of the action of the verb in participial form is the subject of the attached clause,"[1] and so the verb must agree with it. To rephrase the sentence we will have to go through three steps:

1. The verb will be tensed.
2. The subject will be made explicit.
3. An appropriate conjunction (such as *while*, *as*, *because* or *after*) will be used. The output will be something like: *As he was wishing to earn more money, he traveled abroad*; *Because he was hardened by the*

[1] Ronald Forrest, *Revision English*, Essex: Longman, 1968, p. 42.

experience, he did not give up; and, *After he had completed his work, the old man went home*.

6.8. Lexical Gaps

Lexical gaps mean "lexical items or strings which have been completely deleted"[1] from the sentence. English, for example, permits the omission of relative pronouns creating a problem for analysis and consequently for generation where targets are not found to agree with their correct controllers. The problem can be shown the following examples proposed by Hutchins[2]:

> The mathematics students sat their examinations.
> The mathematics students study today is very complex.

The sequence *The mathematics students* must be analyzed differently in each of the above sentences. While in the first sentence it is an NP, in the second sentence it is an NP followed by a relative pronoun (*which*), which is optionally omitted, and then another NP that constitutes the subject of the embedded sentence which modifies the first NP. These structural differences cannot be easily detected by an MT system.

Test Examples: Lexical Gaps

a. Medicine, which students study today, is great.
الطب، الذي طلاب يَدْرسونَ اليوم، عظيمُ.
aṭ-ṭibbu, ʾal-laḏī ṭullāb yadrisūna-l-yawma, ʿaẓīmu.

b. Medicine students study today is great.
طلاب الطبِ يَدْرسونَ اليوم عظماء.
ṭullābu-ṭ-ṭibbi yadrisūna-l-yawma ʿuẓamāʾ.

c. The woman whom we saw yesterday is poor.
إنّ الإمرأةَ التي رَأينَا أمس فقيرةُ.
ʾinna-l-ʾimraʾata-l-latī raʾainā ʾams faqīratu.

d. The woman we saw yesterday is poor.
إنّ الإمرأةَ رَأينَا أمس فقراء.
ʾinna-l-ʾimraʾata raʾainā ʾams fuqarāʾ.

e. The woman I work for is very rich.
الإمرأة اَعْملُ لغنيُ جداً.

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 34.
[2] From W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 90.

ʾal-ʾimraʾatu ʾaʿmalu l-ġaniyyun ǧiddan.

f. The school I went to yesterday is old.
إنّ المدرسة ذهبتُ إلى أمس قديمُ.
ʾinna-l-madrasah ḏahabtu ʾilā ʾams qadīm.

g. The car I told you about is expensive.
السّيارة أنا أخبرتكَ عن غاليُ.
ʾas-sayyāratu ʾanā ʾaḫbartuka ʿan ġālī.

h. The man I depend on has died.
الرّجل أعتمدُ على قَدْ مُتُّ.
ʾar-raǧulu ʾaʿtamidu ʿalā qad mutt.

Test Analysis:

The presence of the relative pronouns in *a* and *c* has enabled the system to make correct agreement between the predicative adjectives and their subjects. The lexical gap in the rest of the examples has led the system to confuse the translation of the embedded sentence as well as of the main sentence, leading to the loss of agreement information.

Sometimes lexical gaps are even more complex when the relative pronoun and the auxiliary verb are missing:

Test Examples: Complex Lexical Gaps

a. The woman carrying her child crossed the road.
الإمرأة تَحْملُ طفلها عَبرَ الطّريقَ.
ʾal-ʾimraʾatu taḥmilu ṭiflahā ʿabara-ṭ-ṭarīq.

b. The man studying engineering failed the exams.
الرّجل يَدْرسُ هَنْدَسَة فَشلتْ الإمتحاناتَ.
ʾar-raǧulu yadrisu handasata fašilati-l-ʾimtiḥānāt.

c. The woman attacked by the terrorists is poor.
هَاجمتْ الإمرأةُ مِن قِبل الإرهابيين فقراء.
hağamati-l-ʾimraʾatu min qibali-l-ʾirhābiyyīna fuqarāʾ.

Test Analysis:

Due to lexical gaps the system confuses the translation and misses agreement information.

Sometimes the infinitive after the particle *to* is omitted. This "occurs when the *to* refers to a verb that has previously been used."[1] For example:

I will come if I'm able to (come).

[1]C. E. Eckersley and J. M. Eckersley, *A Comprehensive English Grammar*, Essex: Longman, 1960, p. 232.

Test Examples: Lexical Gaps: Infinitives

a. I will help you if I'm able to.
أنا سَأُساعدكَ إذا أنا قادرُ إلى.
ʾanā saʾusāʿiduka ʾiḏā ʾanā qādirun ʾilā.

b. You can come if you want to.
أنتَ يُمكنُ أنْ تَجيءَ إذا تُريدَ إلى.
ʾanta yumkinu ʾan taǧīʾa ʾiḏā turīdu ʾilā.

c. He works more than he needs to.
يَعْملُ أكثر من يَحتاجُ إلى.
yaʿmalu ʾakṯar min yaḥtāǧu ʾilā.

Test Analysis:
The translation is faulty because lexical gaps were not handled in the output sentences.

6.8.1. Proposed Solution

The general rule with lexical gaps is that they must be appropriately filled during translation. If it is only the relative pronoun that is missing, an appropriate relative pronoun should be presented explicitly in the target language. The relative pronoun must agree with its controller in number and gender. If both the relative pronoun and auxiliary are missing, the "syntactic generation involves the selection of an appropriate relative pronoun and an appropriately tensed verb form."[1]

There is no equivalent in Arabic to the English relative pronoun (*whom, which, that*) that can serve as an object or object of preposition. Arabic translation contains a relative pronoun and another pronoun that serves as an object or object of preposition.

ʾar-raǧulu-l-laḏī ḏahabtu ʾilai-h
The man who I went to him
The man whom I went to

Regarding the third problem where the infinitive is omitted after the particle *to*, the syntactic generation must provide an explicit verb that agrees in number and gender with the controller.

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 133.

6.9. Intersentential Agreement

Analysis in MT is undertaken at the sentential level. “The majority of MT systems are restricted to the sequential analysis of single sentences.”[1] One of the shortcomings of this approach is the lack of coherence across sentences. In the output translation, it is common to see pronouns which do not agree with their antecedents in number or gender. This is due to the fact that they are not initially co-indexed in the analysis phase. For example, the “meaning or reference of a pronoun is generally evident from its context.”[2] This context is not necessarily provided in the same sentence, but it might be mentioned earlier as the “anaphoric relation may cross sentence boundaries”[3]. The context is essential both for recognizing the reference of pronouns as well as for detecting the intended meaning of words.

Test Examples: Intersentential Agreement

a. I have two sons. They are very diligent.
أنا عِنْدَي إبنان. هم مجتهدون جداً.
ʾanā ʿindī ʾibnān. hum muǧtahidūna ǧiddan.

b. I have two daughters. They are very diligent.
أنا عِنْدَي بنتان. هم مجتهدون جداً.
ʾanā ʿindī bintān. hum muǧtahidūna ǧiddan.

c. I have three sons. They are very diligent.
أنا عِنْدَي ثلاثة أبناءُ. هم مجتهدون جداً.
ʾanā ʿindī ṯalāṯatu ʾabnāʾ. hum muǧtahidūna ǧiddan.

d. I have three daughters. They are very diligent.
أنا عِنْدَي ثلاث بناتُ. هم مجتهدون جداً.
ʾanā ʿindī ṯalāṯu banāt. hum muǧtahidūna ǧiddan.

e. I have three cats. They are very beautiful.
أنا عِنْدَي ثلاث قططُ. هم جميلون جداً.
ʾanā ʿindī ṯalāṯu qiṭaṭ. hum muǧtahidūna ǧiddan.

f. I have a dog. It is black.
أنا عِنْدَي كلبُ. إنه أسود.
ʾanā ʿindī kalb. ʾinnahu ʾaswad.

g. I have a cat. It is black.

---

[1] Ibid., p. 88.
[2] Perrin Smith Corder, *Handbook of Current English*, 3rd ed., Glenview, Illinois: Scott, Foresman and Company, 1968, p. 89.
[3] Rodney Huddleston, *An Introduction to English Transformational Syntax*, Essex: Longman, 1976, p. 251.

أنا عِنْدَي قطةٌ. إنه أسود.
ʾanā ʿindī qiṭah. ʾinnahu ʾaswad.

h. I have three dogs. They are black.
أنا عِنْدَي ثلاثة كلابُ. هم سود.
ʾanā ʿindī ṯalāṯatu kilāb. hum sūd.

i. I have three cats. They are black.
أنا عِنْدَي ثلاث قططُ. هم سود.
ʾanā ʿindī ṯalāṯu qiṭaṭ. hum sūd.

j. Three women escaped from prison. A few days later, the police found the runaway prisoners and arrested them.
ثلاث نِساء هَربنَ من السّجن. بضعة أيام بعد ذلك، الشرطة وَجدتْ السّجناءَ الهاربين واعتقلتهم.
ṯalāṯu nisāʾin harabna mina-s-siğn. biḍʿatu ʾayyāmin baʿda ḏālika, ʾaš-šurṭatu wağadati-s-suğanāʾa-l-hāribīna wa ʾiʿtaqalathu.

k. He was walking beside the river. Many lovely trees were surrounding its banks.
هو كَانَ يَمْشي بجانب النّهر. العديد من الأشجار الجميلةِ كَانتْ تُحيطُ مصارفها.
huwa kāna yamšī bi-ğānibi-n-nahr. ʾal-ʿadīd mina-l-ʾašğāri-l-ğamīlati kānat tuḥīṭu maṣārifahā.

Test Analysis:

Each sentence is translated independently of previous sentences, leading to the loss of information needed for resolving anaphoric and lexical ambiguity.

6.9.1. Proposed Solution

MT systems must accommodate features with which sentences are linked, especially the reference of pronouns and referring expressions. Each sentence cannot be properly interpreted if it is taken as a standalone entity. The information conveyed previously in the text is termed 'contextual knowledge'. Contextual knowledge means "storing text-driven knowledge"[1] that has been mentioned in previous sentences or paragraphs to be used in solving ambiguities. However, one difficulty facing this approach is that it would not be clear what "pieces of knowledge were likely to be used later, or how long they should be stored on the off-chance that they would be needed" [2].

[1] W. J. Hutchins and Harold L. Somers, *An Introduction to Machine Translation*, London: Academic Press, 1992, p. 92.
[2] Ibid., p. 92.

6.10. Distance between Target and Controller

Sometime violation of agreement rules occurs as a result of faulty analysis when there is increased distance between a target and its controller. The subject and verb, for example, can be separated by relative pronouns which can be either mentioned explicitly or omitted.

Test Examples: Distance by Relative Pronouns

a. The young handsome woman speaks loudly.
تَتكلّمُ الإمرأةُ الوسيمةُ الصّغيرةُ بصوت عالي.
tatakallamu-l-ʾimraʾatu-l-wasīmatu-ṣ-ṣaġīratu bi-ṣawtin ʿālī.

b. The young handsome woman attending the lecture speaks loudly.
الإمرأة الوسيمة الصّغيرة تَحْضرُ المحاضرة تَتكلّمُ بصوت عالي.
ʾal-ʾimraʾatu-l-wasīmatu-ṣ-ṣaġīratu taḥḍaru-l-muḥāḍaratu tatakallamu bi-ṣawtin ʿālī.

c. The young handsome woman attending the lecture in the International Conference Center speaks loudly.
الإمرأة الوسيمة الصّغيرة تَحْضرُ المحاضرة في مركزِ المؤتمرِ الدّولي يَتكلّمُ بصوت عالي.
ʾal-ʾimraʾatu-l-wasīmatu-ṣ-ṣaġīratu taḥḍaru-l-muḥāḍarata fī markazi-l-muʾtamari-d-dawliyyi yatakallamu bi-ṣawtin ʿālī.

d. There are poor people who sleep in streets and suffer from misery.
هناك الناسُ الفقراء الذين يَنَامونَ في الشّوارع ويَعانونَ من البؤس.
hunāka-n-nāsu-l-fuqarāʾu-l-laḏīna yanāmūna fī-š-šawāriʿi wa yuʿānūna mina-l-buʾs.

e. There are poor people without homes, who sleep in streets and suffer from misery.
هناك الناسُ الفقراء بدون بيوتُ، التي تَنَامُ في الشّوارع وتَعاني من البؤس.
hunāka-n-nāsu-l-fuqarāʾu bidūn biyūtu, ʾal-latī tanāmu fī-š-šawāriʿi wa tuʿānī mina-l-buʾs.

Test Analysis:
In examples *c* and *e* the system fails to preserve agreement information because of the increased distance between the targets and their controllers.

Appositives can also pose a similar problem. Appositives are made by introducing "predicate nominals and predicate adverbs of location from a constituent into a noun-modifying position of a matrix sentence."[1] They always follow the noun they modify.

[1] Owen Thomas, *Transformational Grammar and the Teacher of English*, New York: Holt, Rinehart and Winston, 1965, p. 95.

Test Examples: Distance by Appositives

a. My aunt loves roses.
عمتي تَحْبُّ وردَ.
ʿammatī tuḥibbu ward.

b. My aunt, a telephone operator, loves roses.
عمتي، عامل بدالة، يَحْبُّ وردَ.
ʿammatī, ʿāmilu baddālah, yuḥibbu ward.

c. His sister came yesterday.
أخته جاءتْ أمس.
ʾuẖtuhu ǧāʾat ʾams.

d. His sister, a computer engineer, came yesterday.
أخته، مهندس حاسبةٍ، جاءَ أمس.
ʾuẖtuhu, muḥandisu ḥāsibah, ǧāʾa ʾams.

e. The girls are diligent.
إنّ البناتِ مجتهداتُ.
ʾinna-l-banāti muǧtahidāt.

f. The girls in this school are diligent.
البنات في هذه المدرسةِ مجتهدةٌ.
ʾal-banātu fī hāḏihi-l-madrasati muǧtahidah.

g. The books are useful.
إنّ الكتبَ مفيدةٌ.
ʾinna-l-kutuba mufīdah.

h. The books in this place are useful.
الكتب في هذا المكان مفيدُ.
ʾal-kutubu fī hāḏā-l-makāni mufīd.

Test Analysis:

In examples *b, d, f,* and *h* the system does not preserve agreement information due to the presence of appositives.

6.10.1. Proposed Solution

Proper analysis, which can deal with long noun phrases and apposition, can be able to avoid confusion in detecting constituents. This enables the generation module to assign agreement features to appropriate elements.

6.11. Agreement Conflict

In some few instances English nouns have gender and are referred to according to that given gender. This leads to confusion when the Arabic equivalent has an opposite gender. For example, the names of countries in English may be feminine or neuter while in Arabic they may be either feminine or masculine.

Test Examples: Agreement Conflict: Gender of Countries

a. Egypt has her own characteristics.
مصر عِنْدَها خصائصها الخاصةُ.
miṣru ʿindahā ẖaṣāʾiṣuhā-l-ẖāṣṣah.

b. Iraq has her own characteristics.
العراق عِنْدَهُ خصائصها الخاصةُ.
ʾal-ʿirāqu ʿindahu ẖaṣāʾiṣuhā-l-ẖāṣṣah.

Test Analysis:
The system is blindly led by the source language and violates rules of agreement of the target language, as shown by example *b*.

The same problem can occur with group nouns. “A group noun refers to a group of people. It can take a singular or plural verb.”[1] When it takes a plural verb, it is referred to with a plural pronoun which creates an agreement confusion during translation.

Test Examples: Agreement Conflict: Group Nouns

a. The army are proud of their victory.
إنّ الجيشَ هو فخور بنصرهم.
ʾinna-l-ǧaiša huwa faẖūrun bi-naṣrihim.

b. The family are proud of their victory.
إنّ العائلةَ هي فخورة بنصرهم.
ʾinna-l-ʿāʾilata hiya faẖūratun bi-naṣrihim.

[1] John Eastwood, *Oxford Practice Grammar*, Oxford: Oxford University Press, 1992, p. 142.

c. The group are proud of their victory.
إنّ المجموعةَ هي فخورة بنصرهم.
ʾinna-l-mağmūʿata hiya faḫūratun bi-naṣrihim.

d. The committee are proud of their victory.
إنّ اللّجنةَ هي فخورة بنصرهم.
ʾinna-l-lağnata hiya faḫūratun bi-naṣrihim.

e. The team are proud of their victory.
إنّ الفريقَ هو فخور بنصرهم.
ʾinna-l-farīqa huwa faḫūrun bi-naṣrihim.

Test Analysis:

The system is also blindly led by the source language and violates rules of agreement of the target language. In all the examples above the possessive pronouns should have been translated as singular to agree with their antecedents.

6.11.1. Proposed Solution

It is the responsibility of the generation module to ensure that targets do not have different gender or number from that of their controllers in accordance with the agreement rules.

Conclusion

Different people take different attitudes towards machine translation (MT). Some people express fear from the project; others doubt its validity; while still others maintain overambition in the project.

Those who fear the project think that the machine will replace the human translator, and in the long run translation as a human creative craft will diminish or come to an end. However, this is not a new attitude. The fear of the machine has risen since the Industrial Revolution. It comes to a peak each time the machine handles a new task which was previously managed completely by humans. Historical evidence proves that machines were never rivals to humans. On the contrary, they constituted marvelous aids. Monotonous, repeated tasks were taken over by the machine while humans have more leisure for doing other creative tasks besides helping and controlling the machine.

Similarly, computers are expected to be of great help for the translators worldwide. The number of translators and their time falls far short of the increasing demand of the translation market. Moreover, most of the translation materials in the market are manuals, periodicals, and scientific journals. This type of materials is characterized by four things. First, it usually has a controlled, well-defined terminology. Second, it has a straightforward style, void of metaphors and flowery language. Third, the frequency of repetition is great; it usually uses the same expressions and even the same sentences. Fourth, it needs speed in translation in order to get to the market on time. In a nutshell, most of the market translation materials are repeated and monotonous, and need to be done in a short time. This is where the machine has been and can be most effective and successful.

When MT systems are successfully launched in the translation workshop, this will give time to translators to focus on the kind of texts that require greater creativity. Translators will also need to acquire new skills. They will need to know how to interact and cooperate with the computer, pre-editing and post-editing texts, updating terminology, and even writing new grammar rules when necessary.

Regarding those who doubt the validity of the whole project of MT and question the ability of the computer to undertake this task, the only answer that can be given to them is the many operational programs available in the market today. Desktop programs as well as Internet

applications provide online translation from and into a large number of world languages. It is true that the MT output is rigid and includes many mistakes. MT in general does not match up to the quality and accuracy of human translation, and this is why it is not yet put to a reliable use. However, the quality of MT output can be improved by many ways: improving system rules and formalisms, controlling the input language to make it clean from flowery and ambiguous words, and post-editing the translation to ensure the accuracy and readability of the output text.

Some people overestimate the prospect of MT and believe that at some point in the future computers will be able to translate from one language into another as easily as they make mathematical operations. However, this idea is erroneous altogether. There is a big difference between mathematics and language. Language is a psychological and sociological phenomenon that has puzzled philosophers, philologists, linguists as well as computer engineers throughout ages. Let me mention only two of the problems faced by MT: firstly a word usually denotes more than one meaning, and there is no clear mapping between words and the intended meaning of the speaker. Secondly, translation requires understanding of the input text, and understanding requires making inductions and inferences and knowledge of the real word, which is to a great extent beyond the ability of the computer. If you can expect a robot to walk, you cannot expect it to race an Olympics athlete. Similarly, you cannot expect the computer to translate all types of text or compete with the quality of human translation.

This research shows that many shortcomings in the output of MT are due to either faulty analysis of the source language text or faulty generation of the target language text. Enhancement to the output can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Fully automated, high quality machine translation (FAHQMT) has not yet been achieved and is not expected to be achieved in the near future. Yet there is a lot that we can do to improve the quality of MT output and increase its usefulness.

The thesis has dealt with one of the features that greatly affect the output of MT, that is agreement. Some languages require agreement among sentence elements more than others. It is only by taking the target language into consideration during analysis of the source language and by properly applying agreement rules during generation that we can improve the quality of translation.

This piece of research concludes that English has only ten agreement features compared to twenty-four features fully utilized in Arabic. Therefore, agreement requirements in Arabic are far more compelling than their English counterparts. Arabic rich morphological variation is due to a great extent to the requisition of agreement features.

The thesis has also identified eleven areas that cause agreement problems in the output in English-Arabic translation. Each area has been fully explained, illustrated with examples, and provided with proposed solutions.

The principal practical achievement of this research is the construction of an agreement test suite. This suite is a collection of examples that can be used in testing different agreement features in English-Arabic translation. These examples have been used in exploring the agreement problems and are grouped together in Appendix I. The applications of a test suite are varied as it can be used in testing, evaluating, and measuring the improvements of any MT system that handles English-Arabic translation.

Bibliography

A Dictionary of Modern Linguistic Terms. (1983) Beirut: Librairie du Liban.

ʾAl-îammādī, Yūsuf, Muḥammad Muḥammad ʾaš-ḥinnāwī and Muḥammad ḥafīq ʿAṭā. (1992) *ʾAl-Qawā ïdu-l-ʾAsāsiyyatu fī-n-Naḥwi wa-ṣ-êarf* [Basic Rules of Syntax and Morphology]. Cairo: General Authority for National Publishing Houses.

Al-Jarf, Reima. (2000) "Grammatical agreement errors in L1/L2 translations." IRAL: *International Review of Applied Linguistics in Language Teaching*. 38.1 (2000): 1-15.

Arnold, Doug, L. Balkan, R. Lee Humphreys, S. Meijer and L. Sadler. (1994) *Machine Translation: An Introductory Guide*. Manchester: Blackwell.

Azar, Betty Schrampfer. (1989) *Understanding and Using English Grammar*. Englewood, New Jersey: Prentice-Hall.

Borsley. Robert. (1999) *Syntactic Theory: A Unified Approach*. 2nd Ed. London: Arnold.

Boulton, Marjorie. (1960) *The Anatomy of Language: Saying what we Mean*. London: Routledge and Kegan Paul Limited.

Brown, Keith and Jim Miller. (1991) *Syntax: A Linguistic Introduction to Sentence Structure*. 2nd Ed. London: Harper Collins Academic.

Catford, J. C. (1965) *A Linguistic Theory of Translation*. Oxford: Oxford University Press.

Chomsky, N. (1981) *Lectures on Government and Binding*. Dordrecht: Foris.

Corder, Perrin Smith. (1968) *Handbook of Current English*. 3rd Ed. Glenview, Illinois: Scott, Foresman and Company.

Cowan, J. Milton. Ed. (1980) *Hans Wehr, A Dictionary of Modern Written Arabic*. 3rd Ed. Beirut: Librairie du Liban.

Eastwood, John. (1992) *Oxford Practice Grammar*. Oxford: Oxford University Press.

Eckersley, C. E. and J. M. Eckersley. (1960) *A Comprehensive English Grammar*. Essex: Longman.

Eynde, Frank Van. Ed. (1993) *Linguistic Issues in Machine Translation*. London: Pinter Publishers.

Fehri, Abdelkader Fassi. (1993) *Issues in the Structure of Arabic Clauses and Words*. Dordrecht: Kluwer Academic Publishers.

Forrest, Ronald. (1968) *Revision English*. Essex: Longman.

Fromkin, Victoria and Robert Rodman. (1974) *An Introduction to Language*. New York: Holt, Rinehart and Winston, Inc.

Greenberg, Hoseph H. Ed. (1978) *Universals of Human Language*. Stanford, California: Stanford University Press.

Hairston, Maxine. (1986) *Contemporary Composition*. 4th Ed. Boston: Houghton Mifflin Company.

Hasan, Abbas. (1999) *ʾAn-Naḥwu-l-Wāfī* [Comprehensive Grammar]. 14th Ed. Cairo, Egypt: Dar Al-Maʿarif.

Hockett, Charles F. (1958) *A Course in Modern Linguistics*. New York: The Macmillan Company.

Horrocks, Geoffrey. (1987) *Generative Grammar*. London: Longman.

Huddleston, Rodney. (1976) *An Introduction to English Transformational Syntax*. Essex: Longman.

Hutchins, W. John. (1986) *Machine Translation: Past, Present, Future*. West Sussex, England: Ellis Horwood Limited.

Hutchins, W. John and Harold L. Somers. (1992) *An Introduction to Machine Translation*. London: Academic Press.

Koutsoudas, Andreas. (1966) *Writing Transformational Grammars: An Introduction*. New York: McGraw-Hill.

Laffling, John. (1991) *Towards High-Precision Machine Translation: Based on Contrastive Textology*. Berlin: Foris Publications.

Langacker, Ronald W. (1973) *Language and its Structure: Some Fundamental Linguistic Concepts*. 2nd Ed. New York: Harcourt Brace Jovanovich.

Lehrberger, John and Laurent Bourbeau. (1988) *Machine Translation: Linguistic characteristics of MT systems and general methodology of evaluation*. Amsterdam: John Benjamins Publishing Company.

Lester, Mark. (1976) *Introductory Transformational Grammar of English*. 2nd Ed. New York: Holt, Rinehart and Winston.

Longman Active Study Dictionary. (1986) Essex: Longman.

Longman Dictionary of English Language and Culture. (1992) Essex, England: Longman.

Lyons, John. Ed. (1970) *New Horizons in Linguistics*. Middlesex, England: Penguin Books Ltd.

MacCallum, George P. (1978) *More Idiom Drills for Students of English as a Second Language*. New York: Harper and Row, Publishers.

Massoud, Mary M. F. (1988) *Translate to Communicate: A Guide for Translators*. Elgin, Illinois: David C. Cook.

Matthews, P. H. (1981) *Syntax*. Cambridge: Cambridge University Press.

Microsoft Bookshelf 98. (1998) CD. Microsoft.

Munn, Alan. (1999) "First Conjunct Agreement: Against a Clausal Analysis." *Linguistic Inquiry*. 30.4, 643-669.

Murphy, Raymond. (1985) *English Grammar in Use*. Cambridge: Cambridge University Press.

Nasr, Raja T. (1967) *The Structure of Arabic: From Sound to Sentence*. Beirut: Librairie du Liban.

Nesfield, J. C. and F. T. Wood. (1964) *Manual of English Grammar and Composition*. London: Macmillan.

Nirenburg, Sergei. Ed. (1987) *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press.

Palmer, Frank. (1971) *Grammar*. Middlesex: Penguin Books.

Popowich, Fred, Paul McFetfidge, Davide Turcato and Janine Toole. (2000) "Machine Translation of Closed Captions." *Machine Translation*. 15, 311-341.

Potter, Simeon. (1950) *Our Language*. Middlesex, England: Penguin.

Quirk, Randolph and Sidney Greenbaum. (1973) *A University Grammar of English*. Harlow, Essex: Longman.

Radford, Andrew. (1988) *Transformational Grammar: A First Course*. Cambridge: Cambridge University Press.

_____. (1997) *Syntax: A Minimalist Introduction*. Cambridge: Cambridge University Press.

Rayner, Manny, David Carter, Pierrette Bouillon, Vassilis Digalakis and Mats Wirén. Eds. (2000) *The Spoken Language Translator*. Cambridge: Cambridge University Press.

Reid, Wallis. (1991) *Verb and Noun Number in English: A Functional Explanation*. London: Longman.

Roach, Peter. Ed. (1992) *Computing in Linguistics and Phonetics: Introductory Readings*. San Diego: Academic Press.

Roberts, Ian. (1997) *Comparative Syntax*. London: Arnold.

Robinson, Douglas. (1997) *Becoming a Translator: An Accelerated Course*. London: Routledge.

Rule, Lareina. (1963) *Name Your Baby*. New York: Bantam Books.

Soltan, Usama Mohamed. (1996) "A Contrastive and Comparative Syntactic Analysis of Deletion Phenomena in English and Standard Arabic" Unpublished MA Dissertation. Ain Shams University.

The American Heritage Dictionary of the English Language. 3rd Ed. (1996) Boston: Houghton Mifflin.

Thomas, Owen. (1965) *Transformational Grammar and the Teacher of English*. New York: Holt, Rinehart and Winston.

Toivonen, Ida. (2000) "The Morphosyntax of Finnish Possessives". *Natural Language and Linguistic Theory*. 18. 579-609.

Trujillo, Arturo. (1999) *Translation Engines: Techniques for Machine Translation*. London: Springer.

Woolford, Ellen. (1999) "More on the Anaphor Agreement Effect." *Linguistic Inquiry*. 30.2, 257-287.

Appendix I: Agreement Test Suite

This is a list of test examples that has been used throughout the thesis to show problematic issues related to agreement. They can be used as a test suite to evaluate the ability of an MT to handle agreement questions. They can also be used during the development process to measure improvements introduced to the system.

(Section 6)

1. A diligent rich handsome man
2. A diligent rich handsome woman
3. Diligent rich handsome men
4. Diligent rich handsome women
5. I saw the diligent rich handsome men.
6. The man who drives the car
7. The woman who drives the car
8. The men who drive the car
9. The women who drive the car
10. This man
11. This woman
12. These men
13. These women
14. That man
15. That woman
16. Those men
17. Those women
18. The boy goes to the garden and waters the flowers.
19. The girl goes to the garden and waters the flowers.
20. The boys go to the garden and water the flowers.
21. The girls go to the garden and water the flowers.
22. The boy went to the garden and watered the flowers.
23. The girl went to the garden and watered the flowers.
24. The boys went to the garden and watered the flowers.
25. The girls went to the garden and watered the flowers.
26. one boy
27. one girl
28. two boys
29. two girls
30. three boys
31. three girls
32. four boys
33. four girls
34. ten boys
35. ten girls
36. eleven boys
37. eleven girls
38. twelve boys
39. twelve girls

40. fifteen boys
41. fifteen girls
42. twenty-one boys
43. twenty-one girls
44. seventy-eight boys
45. seventy-eight girls

(Section 6.1)
46. You are a good boy.
47. You are two good boys.
48. You are good boys.
49. You are good.
50. You are a good girl.
51. You are two good girls.
52. You are good girls.
53. They are two good boys.
54. They are good boys.
55. They are good cats.
56. They are two good girls.
57. They are good girls.
58. It is a good bull.
59. It is a good cow.
60. The two boys said, "We are good."
61. The two girls said, "We are good."
62. The boys said, "We are good."
63. The girls said, "We are good."
64. The boy said, "I am good."
65. The girl said, "I am good."

(Section 6.2)
66. Bush read the book.
67. Harmony read the book.
68. Carpenter read the book.
69. Hope read the book.
70. Foot read the book.
71. Flora read the book.
72. Lance read the book.
73. Aura read the book.
74. Jack read the book.
75. Jane read the book.
76. Jackson read the book.
77. Janet read the book.
78. Nicolas read the book.
79. Josephine read the book.
80. Pierre read the book.
81. Simone read the book.
82. Vasco read the book.
83. Isabella read the book.
84. Antonio read the book.
85. Maria read the book.

86. Ahmad read the book.
87. Salma read the book.
88. Hassan read the book.
89. Nada read the book.

(Section 6.3)
90. She is a good singer.
91. These girls are students.
92. These women are engineers.
93. These women are teachers.
94. These girls are good players.
95. These women are good actors.
96. The student likes her teacher.
97. The students like their teachers.

(Section 6.4)
98. I held him with my hands.
99. I saw him with my eyes.
100. I heard him with my ears.
101. My legs cannot carry me.
102. These two men
103. These two women
104. Those two men
105. Those two women
106. The two men who smile
107. The two women who smile
108. They are two good boys.
109. They are two good girls.
110. Two active, diligent women attended the meeting.
111. Two active, diligent men attended the meeting.
112. The two girls love their mother.
113. Both boys love football.
114. Both girls love football.
115. Both of them love football.
116. The boy and the girl are happy.
117. She and I were school fellows.
118. He and his wife are always fighting.
119. Jack and Jane went to the garden and played football.
120. Jack and John went to the garden and played football.
121. He finished the work after a couple of hours.
122. A couple of boys were playing in the garden.
123. A couple of girls were playing in the garden.
124. A couple of birds were flying over there.
125. He drank a couple of glasses.

(Section 6.5)
126. The boy must depend on himself.
127. The girl must depend on herself.
128. We must depend on ourselves.
129. The man said to his son, "You must depend on yourself."

130. The man said to his daughter, "You must depend on yourself."
131. The man said to his two sons, "You must depend on yourselves."
132. The man said to his sons, "You must depend on yourselves."
133. The man said to his daughters, "You must depend on yourselves."
134. The two boys must depend on themselves.
135. The two girls must depend on themselves.
136. The boys must depend on themselves.
137. The girls must depend on themselves.
138. The cats must depend on themselves.
139. The cow must depend on itself.
140. The bull must depend on itself.
141. The two boys love each other.
142. The two girls love each other.
143. The boys love each other.
144. The girls love each other.
145. The boys believe that they are diligent.
146. The two boys believe that they are diligent.
147. The two girls believe that they are diligent.
148. The girls believe that they are diligent.
149. The girls went to bed because they were tired.
150. The girls met their brother while they were walking in the garden.
151. The girls met their brothers while they were walking in the garden.
152. The boy depends on his father.
153. The girl depends on her father.
154. The boys depend on their father.
155. The two boys depend on their father.
156. The two girls depend on their father.
157. The girls depend on their father.
158. The girls and their brother are diligent.
159. The cats depend on their father.
160. The bull depends on its father.
161. The cow depends on its father.
162. The book and its cover were torn.
163. The table and its cover were cleaned.
164. They depend on the boy.
165. The boy depends on their father.
166. The boy depends on them.

(Section 6.5.1)

167. The men met the women and they were happy.

(Section 6.6.1)

168. The boy seems to be happy.
169. The girl seems to be happy.
170. The boys turned out to be smart.
171. The girls turned out to be smart.
172. The boy is believed to be good.
173. The girl is believed to be good.
174. The boys appear to be good.
175. The girls appear to be good.

176. The boys happened to be smart.
177. The girls happened to be smart.
178. The girl is likely to be happy.
179. The girls are certain to be happy.

(Section 6.6.2)

180. He persuaded her to visit his father.
181. He wanted to visit his father.
182. He wanted her to visit his father.
183. He appealed to her to visit his father.
184. She went to England to learn English.
185. She is able to go to the party.
186. He promised her to visit his father.
187. My aim was to make money.
188. His aim was to make money.
189. Her aim was to make money.
190. The task of the manager is to organize work.
191. He thought that his responsibility was to bring peace to the region.
192. We are not in a hurry to know the results.
193. They have nothing to do.
194. She found nothing to eat.
195. It was silly of them to come late.
196. To quit the job now will be wrong.
197. There was nothing to eat in the fridge.

(Section 6.6.3)

198. He can speak English.
199. She can speak English.
200. The boys could understand him.
201. The girls could understand him.
202. The boys should help him.
203. The girls should help him.
204. I will help you.
205. You must go now.
206. I saw the boy carry the book.
207. I saw the girl carry the book.
208. They heard her speak.
209. They heard him speak.
210. Don't make me laugh.
211. He let the boys use his car.
212. He let the girls use his car.
213. I can't avoid meeting him.
214. I will consider visiting him.
215. He denied attacking her.
216. I can't risk losing that job.

(Section 6.7)

217. Wishing to earn more money, he traveled abroad.
218. Going to school, the student lost his bag.
219. Born in England, she spoke perfect English.

220. Hardened by the experience, he did not give up.
221. Taken by surprise, she gave up.
222. Having completed his work, the old man went home.
223. Having completed her work, the old woman went home.
224. Having been abroad for thirty years, they knew many languages.
225. Having been abroad for thirty years, she knew many languages.

(Section 6.8)

226. Medicine, which students study today, is great.
227. Medicine students study today is great.
228. The woman whom we saw yesterday is poor.
229. The woman we saw yesterday is poor.
230. The woman I work for is very rich.
231. The school I went to yesterday is old.
232. The car I told you about is expensive.
233. The man I depend on has died.
234. The woman carrying her child crossed the road.
235. The man studying engineering failed the exams.
236. The woman attacked by the terrorists is poor.
237. I will help you if I'm able to.
238. You can come if you want to.
239. He works more than he needs to.

(Section 6.9)

240. I have two sons. They are very diligent.
241. I have two daughters. They are very diligent.
242. I have three sons. They are very diligent.
243. I have three daughters. They are very diligent.
244. I have three cats. They are very beautiful.
245. I have a dog. It is black.
246. I have a cat. It is black.
247. I have three dogs. They are black.
248. I have three cats. They are black.
249. Three women escaped from prison. A few days later, the police found the runaway prisoners and arrested them.
250. He was walking beside the river. Many lovely trees were surrounding its banks.

(Section 6.10)

251. The young handsome woman speaks loudly.
252. The young handsome woman attending the lecture speaks loudly.
253. The young handsome woman attending the lecture in the International Conference Center speaks loudly.
254. There are poor people who sleep in streets and suffer from misery.
255. There are poor people without homes, who sleep in streets and suffer from misery.
256. My aunt loves roses.
257. My aunt, a telephone operator, loves roses.
258. His sister came yesterday.
259. His sister, a computer engineer, came yesterday.

260. The girls are diligent.
261. The girls in this school are diligent.
262. The books are useful.
263. The books in this place are useful.

(Section 6.11)

264. Egypt has her own characteristics.
265. Iraq has her own characteristics.
266. The army are proud of their victory.
267. The family are proud of their victory.
268. The group are proud of their victory.
269. The committee are proud of their victory.
270. The team are proud of their victory.

Appendix II: Al-Mutarğim Al-Arabey Demo CD