Intelligent Information Management, 2012, 4, 13-22
http://dx.doi.org/10.4236/iim.2012.41003 Published Online January 2012 (http://www.SciRP.org/journal/iim)

An English-to-Arabic Prototype Machine Translator for Statistical Sentences

Hamdy N. Agiza[1], Ahmed E. Hassan[2], Noura Salah[1]
[1]Mathematics Department, Faculty of Science, Mansoura University, Mansoura, Egypt
[2]Electrical Engineering Department, Faculty of Engineering, Mansoura University, Mansoura, Egypt
Email: agiza@mans.edu.eg, arwaahmed1@gmail.com, noura_salah91@yahoo.com

Received July 9, 2011; revised September 14, 2011; accepted September 25, 2011

ABSTRACT

Authors of that paper proposed a prototype machine translator system to translate scientific English sentences into Arabic sentences. This system is based on natural language processing and machine learning. This proposed system is applied in statistical field, which is very important on a mathematical sub field in Math department. The system is analyzed, designed and developed. Author tested the proposed system on some statistical statements. It proves its validity as a prototype system.

Keywords: Machine Translation; Natural Language Processing; Parsing; Dictionary

1. Introduction

Machine Translation (MT) is one of the applications of Natural Language Processing (NLP) [1]. Also it is called Automatic Translation and it has been defined as: “the process that utilizes computer software to translate text from one natural language to another” [2]. This definition involves the semantic mapping from one natural language to another. MT is therefore defined as “the transfer of meaning from one (human) language to another with the aid of a computer” Goshawke al 1987 [1]. Also it is accounting for the grammatical structure of each language and using rules and assumptions to transfer the grammatical structure of the source language (text to be translated) into the target language (translated text).

This definition stresses the fact that MT is not simply substituting words for other words, but like human it involves the application of complex linguistic rules especially in morphology, syntax and semantics. Which means that computer could be used to translate from source language to target language. It could translate an entire document automatically and then presents it to a human. But when a human composes a translation, perhaps calling on a computer for assistance in specific tasks such as looking up specialized words and expressions in a dictionary, the process is called human translation. There is a gray area between human and machine translation, in which the computer may retrieve whole sentences of previously translated text and make minor adjustments as needed. However, even in this gray area, each sentence was originally the result of either human translation or machine translation. Authors should reserve the label MT for the case when a computer performs both the initial translations of the sentences and subsequent manipulations, which could be called translator tools.

There are three basic approaches being used for developing MT systems [4] which differ in their complexity and sophistication. These approaches are: [2,4] direct or transformer approach, transfer-based approach, and Interlingua approach. The proposed system should follow direct MT approach, explained in details in Section 3.

2. Review

The idea of using computers to translate or help translate human languages is almost as old as the computer itself. Before email, before word processing, before command-line interfaces, machine translation or MT—was one of the first two computer applications designed to act upon words instead of numbers (the other was code breaking). But it turns out that really good MT is so hard to pull off that the task exhausted the top-end computing resources of every generation attempting it. Regardless, machine translation is going stronger than ever, fired up by the globalization of the Net. Today, all over the world, software designers, programmers, hardware engineers, neural-network experts, AI specialists, linguists, and cognitive scientists are enlisted in the effort to teach computers how to port words and ideas from language to language [9].

New developments worldwide in the fields of technology, political and socioeconomic trends starting in the

1980s contributed to a revival of Machine Translation advancement and research. These developments include the strides made in information technology, a rapid fall in the cost of computing power, globalization and increaseing demand from multinational companies and governments for translation. These developments are by no means the prime mover of research and development behind MT; they just helped increase the pace of development. Translation and Interpretation asserts that research and development of MT has been going on since the 1950s "engaging some of the best minds in computing, linguistics and artificial intelligence," [9].

As English is a universal language, most of the researches in MT are mainly concentrated on the translation between English and Arabic because automatic English-to-Arabic translation is still an active area and this will help in simplifying the Arab communication with other countries. There were a lot of implemented systems that work especially in English-to-Arabic MT field:

Hoda M.O. Mokhtar *et al.* [5] have proposed MT system, which is an automatic system for English-to-Arabic translation of scientific text. (SEATS) adopts a transfer approach that employs a unification based grammar, transformation rules and Arabic morphological synthesis rules.

Beesly [7], it is a finite-state morphological analyzer for Arabic words which is consists of analyzer proper, running on a network server, and Java applets that run on the user's machine and render words in standard Arabic orthography both for input and output.

Al-Anzi *et al.* [3] have proposed MT system to translate English web pages to Arabic. The system partitions the English sentence into different parts according to HTML tag occurs. Then it translates the part of the English sentence independently of others and inserts the translation between the HTML tags that were present in the source. Its result showed that the system had faced difficulties when an HTML tag appeared inside a sentence [1].

There are commercial MT systems. "Al-Mutarjim Al-Arabey" which translates English text into Arabic [1,13], "golden Al-Wafi translator" which also translates English text into Arabic [12] and "Sakhr CAT" translator is a computer-aided translation system supporting bidirectional bilingual translation between English and Arabic [5].

The previous lectures proposed a generic MT system but domain specific MT systems are mandatory for scientific applications. Martha Palmer *et al.* [8] proposed a prototype system applied on domain specific MT which translates the military text. Rafea *et al.* (1992) developed an English-Arabic MT system which translates a sentence from the domain of the political news of the Middle East. Pease *et al.* (1996) developed a system which translates medical texts from English to Arabic. Mokhtar (2000) developed an English-Arabic MT system which applied to abstracts from the field of Artificial Intelligence [6]. Authors of that paper proposed MT system for domain specific (the statistical sentences).

The translation in Arabic language still limited and its results are still not totally satisfactory [5]. Little work has been done in developing Arabic-to-English MT systems. Al Barhamtoshy (1995) proposes a translation method for compound verbs. Shaalan (2000) described a tool for translating the Arabic interrogative sentence into English. Chalabi (2001) presented an Arabic-English MT engine that allows any Arabic user to search and navigate through the Internet using the Arabic language. Othman *et al.* (2003) developed an efficient chart parser that will be used for translating Arabic sentence [6].

3. Problem Definition

Although English is a universal language and most of the researches in MT are mainly concentrated on the translation between English and Arabic language. But a few translations in scientific fields are existing. Authors specify a statistical field as a specific domain in the proposed system translation. This field is very important on a mathematical sub field in Mathematics department, which ease the teaching and research in that department.

4. The Proposed MT System

The main idea behind the proposed MT system is to translate the Source Language (SL) sentences to the Target Language (TL) sentences by carrying out the possible parse, replacing source words with their target language equivalents as specified in a bilingual dictionary, and then re-arranging their order to suit the rules of the target language.

4.1. Abstract Architecture of the System

The first stage of processing involves the parser [4], which does some preliminary analysis of the source sentence. This is passed to a package of rules which transform the sentence into a target sentence, using necessary information provided by the parsing process. The transformation rules include bilingual dictionary rules and various rules to reorder words as shown in **Figure 1**. The transformer system is really designed with translation in one direction, between one pair of languages in mind.

As English is a universal language, most of the researches in Arabic MT are mainly concentrated on the translation between English and Arabic [6, 1, and 5]. This could help in simplifying the Arab communication with other countries.

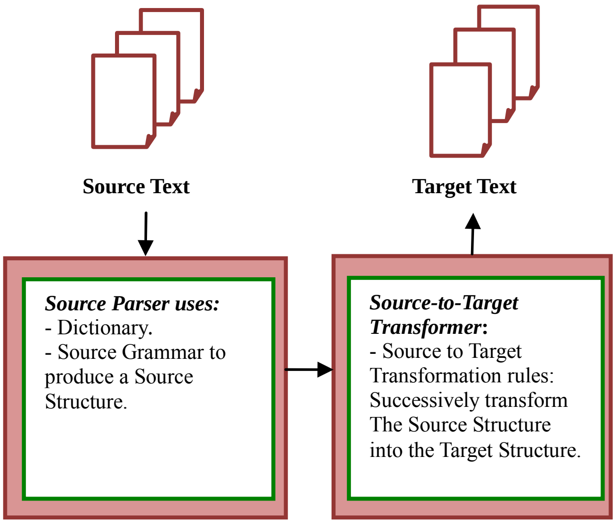

Source Text
Target Text
Source Parser uses:
- Dictionary.
- Source Grammar to produce a Source Structure.
Source-to-Target Transformer:
- Source to Target Transformation rules: Successively transform The Source Structure into the Target Structure.

Figure 1. Components of MT proposed system.

4.2. Proposed System Detailed Architecture

The current system is proposed for scientific English-Arabic automatic translation. It consists of three main modules as shown in **Figure 2**:

- Analysis_module: This is used to analyze the input text.
- A transformer module: This is used to translate English sentences structures and words.
- A generation module: This is used to produce target Arabic sentences behind its input-output interface and special MT requirements.

The proposed system architecture is shown in **Figure 2**. The current system is developed in visual c#.Net programming language environment and SQL server as the associated database management system.

4.2.1. Special MT Requirements

The proposed system centered on the domain-specific data founding in an English-Arabic Bilingual Dictionary selecting from statistical books in which Efficient MT system would rely heavily on domain-specific statistical vocabulary. This nature is not supported by most English-to-Arabic translators systems. For any specific MT we would have to augment extensively with additional domain-specific vocabulary.

The proposed system composed of three main components. As described in the previous section the first is used to analyze the input English sentence and uses Dictionary and suitable Grammar to Produce a Source (English) Structure. Secondly is a transformer module which is used to translate Source (English) Structure and words to target (Arabic) language structure and words. And finally a generation module which is used to produce target (Arabic) language text. The flow of all the process is shown in **Figure 3** which the proposed system flow chart.

4.3.2. Analysis Module

The analysis is done throw two main phases: scanner and parser phases.

4.3.2.1. Scanner

The English sentence entered to the proposed system. And then it uses the Scanner, which it divides the English sentence into words by splitting it when it finds a space string. The output of this step is a list of English words that ready to go to parser. Author called the output list “English_words_list”, assigns the number of words in the sentence to the variable name “English_words_list_lentgh” and should keep the order of words as shown in **Figure 4**.

4.3.2.2. Parser

English sentences Analysis (Parsing), it means that: employing the possible English grammar rules that have been selected to cover almost sentences cases that could compose an abstract encountered by the system to analyze (parse) the input English sentence and Produce an English Structure. Rules were implemented using Phrase Structure Grammar (PSG) [5]. English sentences analysis is carried out through bottom-up parser.

The parser accepts “English_words_list “that building a sentence and output a list of parts of speech like noun, verb, determinant, auxiliaries, adjective, preposition and etc as shown in **Figure 5**. It adopts semantic features to finally accept sentences that are grammatically and semantically correct. Authors used the miniature English grammar and lexicon table as shown in [11]. The following example explain how it works. The input sentence is:

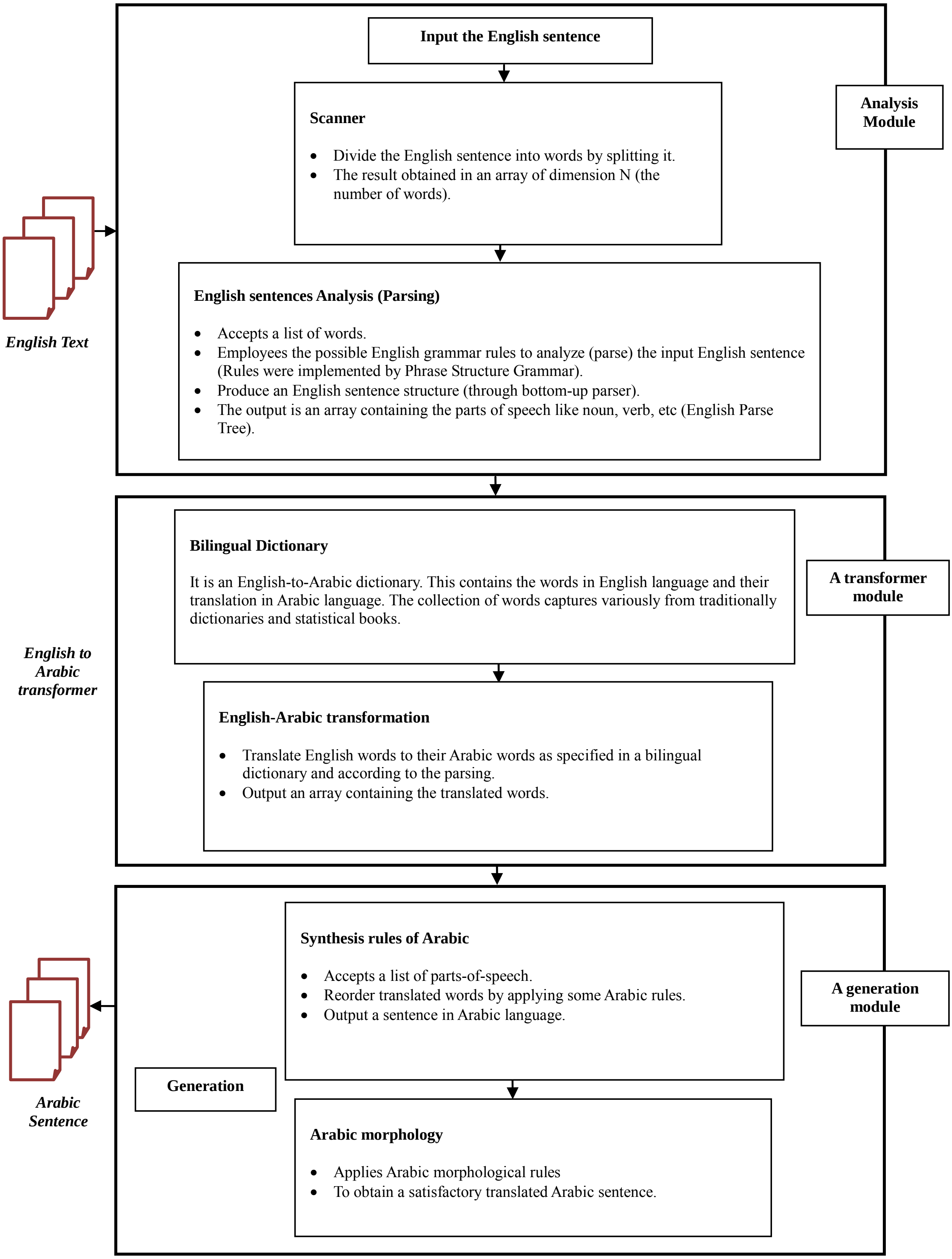
The statistics is the analysis of the data.

To parse this sentence which it has [the, statistics, is, the, analysis, of, the, data] English_words_list, linguists often use a special notation to write out grammar rules. In this notation, a rule consists of a “left-hand-side” (LHS) and a “right-hand-side” (RHS) connected by an arrow (→) [10] as shown below:

S → NP VP; VP → V NP PP; PP → preposition NP; NP → (DET) N; DET→ the; N → statistics; V → is; DET → the; N → analysis; P → of; DET → the; N → data.

Where S means sentence, NP means noun phrase, VP means verb phrase, DET means determinant *or article like* “*the*, *a and an*”, N means noun, V means verb and P means preposition.

Semantics is concerned with the meaning of words and how they combine to form sentence meanings; there are many ways of thinking about representing word meanings, it involves associating words with semantic features which correspond to their sense components. Associating



Input the English sentence

Scanner

- Divide the English sentence into words by splitting it.
- The result obtained in an array of dimension N (the number of words).

English sentences Analysis (Parsing)

- Accepts a list of words.
- Employees the possible English grammar rules to analyze (parse) the input English sentence (Rules were implemented by Phrase Structure Grammar).
- Produce an English sentence structure (through bottom-up parser).
- The output is an array containing the parts of speech like noun, verb, etc (English Parse Tree).

Bilingual Dictionary

It is an English-to-Arabic dictionary. This contains the words in English language and their translation in Arabic language. The collection of words captures variously from traditionally dictionaries and statistical books.

English-Arabic transformation

- Translate English words to their Arabic words as specified in a bilingual dictionary and according to the parsing.
- Output an array containing the translated words.

Synthesis rules of Arabic

- Accepts a list of parts-of-speech.
- Reorder translated words by applying some Arabic rules.
- Output a sentence in Arabic language.

Arabic morphology

- Applies Arabic morphological rules
- To obtain a satisfactory translated Arabic sentence.

Figure 2. Architecture of the proposed system.

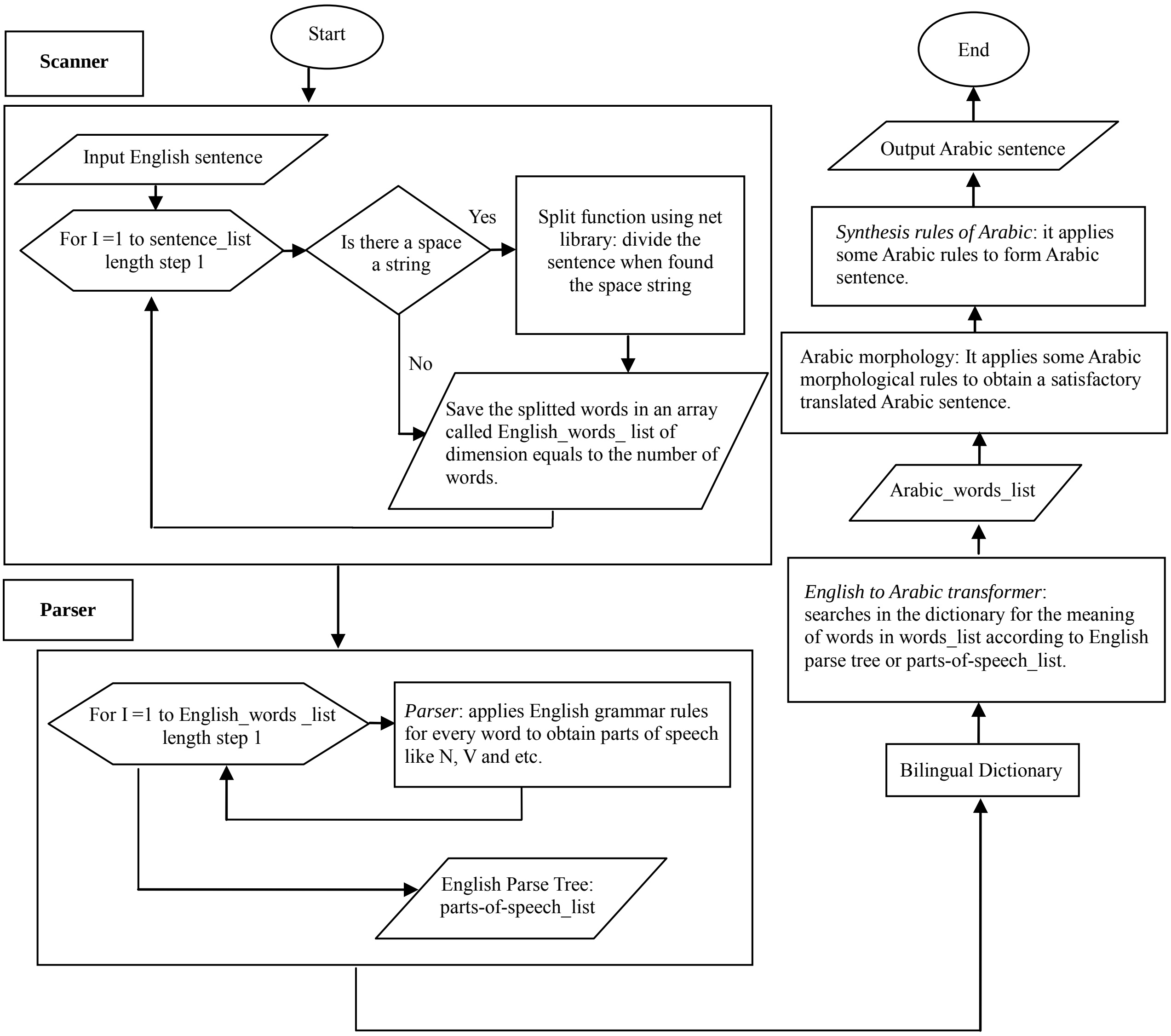

Scanner
Start
Input English sentence
For I =1 to sentence_list length step 1
Is there a space a string
Yes
No
Split function using net library: divide the sentence when found the space string
Save the splitted words in an array called English_words_ list of dimension equals to the number of words.
Parser
For I =1 to English_words _list length step 1
Parser: applies English grammar rules for every word to obtain parts of speech like N, V and etc.
English Parse Tree: parts-of-speech_list
Bilingual Dictionary
English to Arabic transformer: searches in the dictionary for the meaning of words in words_list according to English parse tree or parts-of-speech_list.
Arabic_words_list
Arabic morphology: It applies some Arabic morphological rules to obtain a satisfactory translated Arabic sentence.
Synthesis rules of Arabic: it applies some Arabic rules to form Arabic sentence.
Output Arabic sentence
End

Figure 3. MT proposed system flow chart.

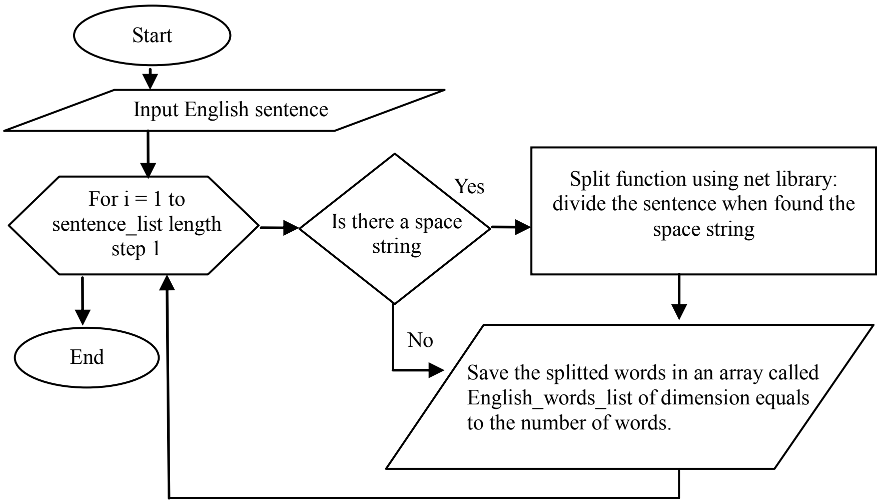

Start
Input English sentence
For i = 1 to sentence_list length step 1
Is there a space string
Yes
No
Split function using net library: divide the sentence when found the space string
Save the splitted words in an array called English_words_list of dimension equals to the number of words.
End

Figure 4. Scanner flow chart.

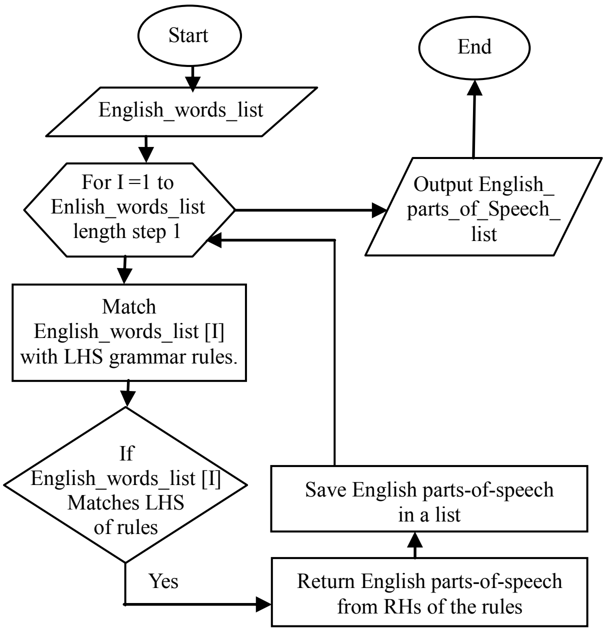


Figure 5. Parser flow chart.

words with semantic features is useful because some words impose semantic constraints on what other kinds of words they can occur with.

After obtaining the English_parts_of_speech_list, some semantic features has been applied for every word in English_words_list, in which it deals with the relation between categories such as "Subject", "Object" and (deep) categories such as "Agent" and "Effect". It reduces the ambiguity of choosing the meaning of words.

4.3.3. Transformer Module

The transformation is done throw two phases: Building a Bilingual dictionary and English-Arabic transformation:

4.3.3.1. Building a Bilingual Dictionary

A Bilingual dictionary is an English-to-Arabic dictionary that contains the words in English language and their translation in Arabic language. Author has been used sql server database as the associated management system. To form tables in the database, author collected words variously from different traditional dictionaries and statistical books to cover the statistical vocabularies that may found in the input sentence. And also it contains the word characters such as type, gender, tense, numbers and meaning.

4.3.3.2. English-to-Arabic Transformation

The module accepts "English_words_list" and "English_ parts-of-speech_list". The output is "Arabic_words_list". The system looks up in the bilingual dictionary for the Translation of English words and obtains equivalent Arabic words Translation according to the transformer flow chart as shown in **Figure 6**.

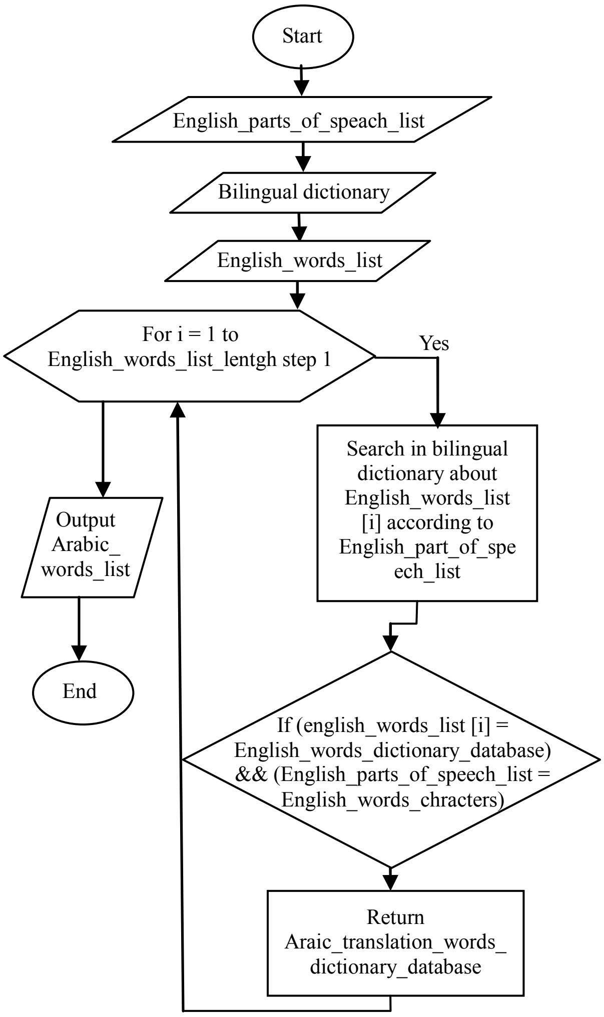


Figure 6. Transformer flow chart.

4.3.4. Generation Module

Generates translated Arabic sentence after applying transformation rules is done within that module through two phases (Synthesis rules of Arabic) and (Synthesis rules of Arabic).

4.3.4.1. Synthesis Rules of Arabic

At that phase the system accepts "*Arabic_words_list*" and the output is a sentence in a target (Arabic) language. It is the previous final phase that reordering translated words according to various Arabic rules as shown in **Figure 7**. The Arabic sentence is generated from English sentence by some of the following rules:

1) Verb phrase in Arabic sentence has the order to form as follow:

a) The subject in English sentence is located after the verb.

b) The object in English sentence is located after the subject.

2) Noun phrase in Arabic sentence has the same order

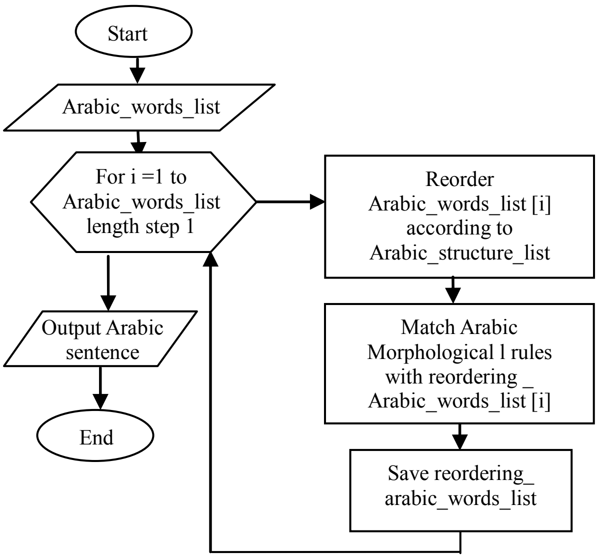

Start
Arabic_words_list
For i =1 to Arabic_words_list length step 1
Reorder Arabic_words_list [i] according to Arabic_structure_list
Match Arabic Morphological l rules with reordering _ Arabic_words_list [i]
Save reordering_ arabic_words_list
Output Arabic sentence
End

Figure 7. Generation module flow hart.

to form like *English* sentence.

3) If the English sentence matches the rule

NP → DET N then DET located before N for example "the book" English sentence would be "ال كتاب" in Arabic sentence.

4) If the English sentence match the rule NP → ADJ N then ADJ located after N for example "clever boy" in the English sentence would be "ولد ماهر" in the Arabic sentence. Where ADJ means adjective.

4.3.4.2. Arabic Morphology

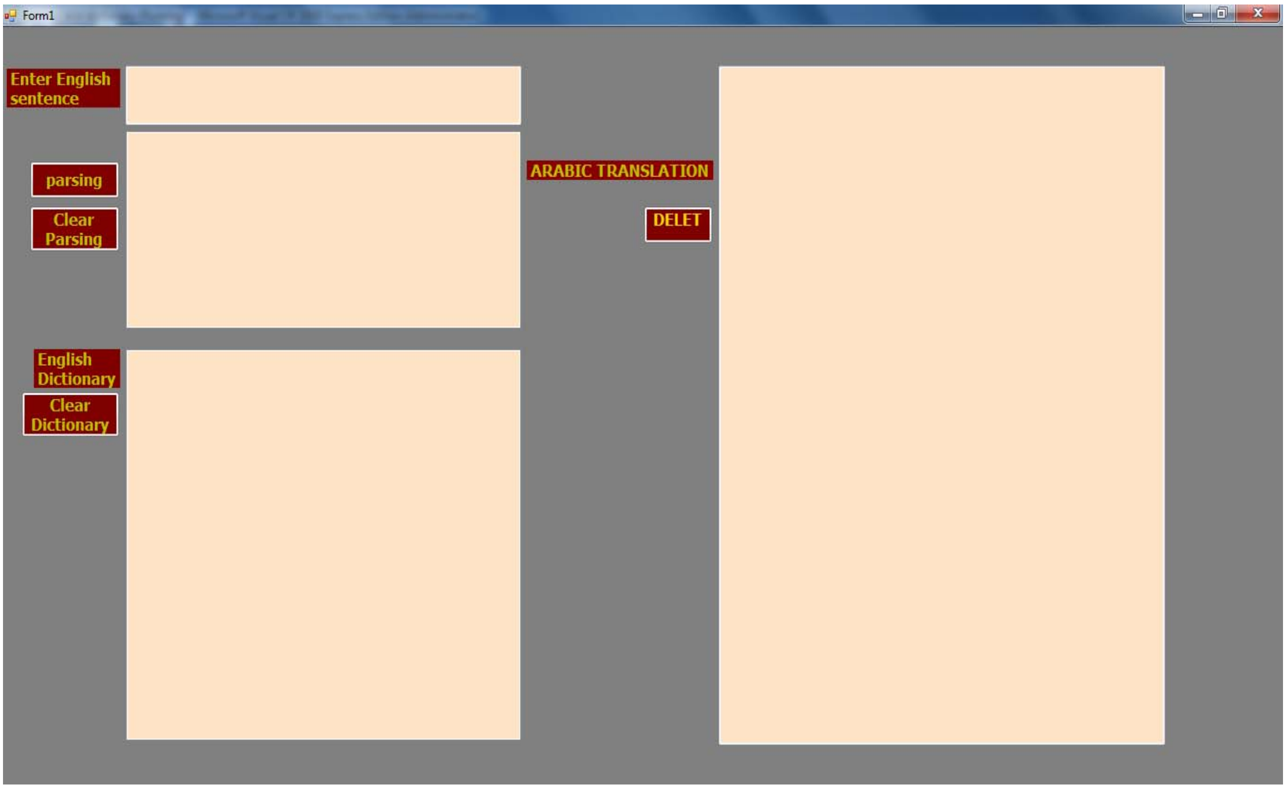
After obtaining translated Arabic sentence, the system has to apply Arabic morphological rules to obtain satisfactory Arabic sentence.

For example "*the girls play in the garden*", here girls are female so the translation of the verb play should be "يلعبن" which it combines of translation of play "يلعب" and morphological rule "ن النسوه" In contradiction that if the sentence be "the boy play in the garden", the translation should be "يلعبون" because boy's gender is male so it combines of play "يلعب" and morphological rule " ون للجمع المذكر".

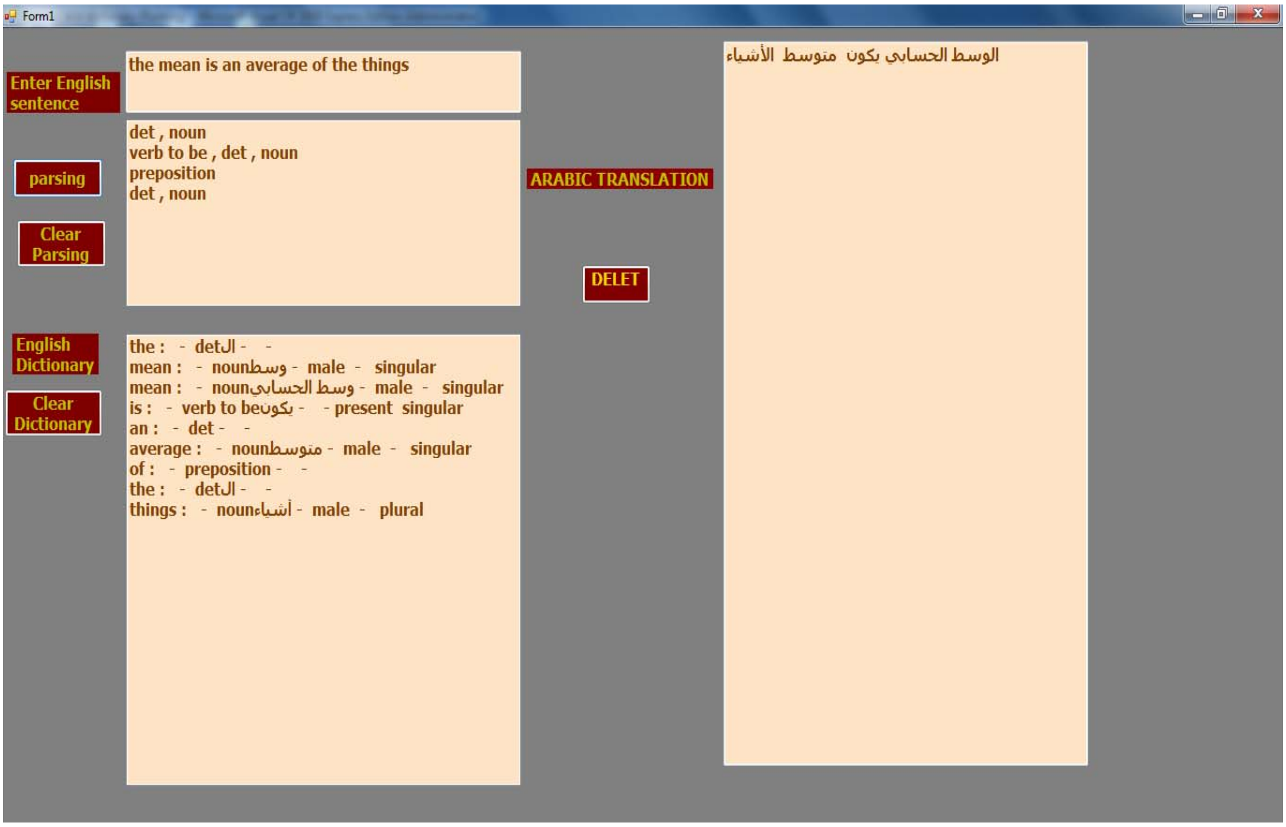
5. System Implementation and Results

The system will be evaluated through performing a number of successful tests by randomly choosing a set of sentences from the field of mathematical studies and that to prove its validity. A dictionary is implemented by choosing statistical vocabularies from statistical books, Elias, Almawred and *et al.* [1].

The proposed system is being carried out in visual c#.Net programming language environment and SQL server as the associated database management system. And some shoots from the results shown below in **Figures 8(a)-(e)**.



(a)



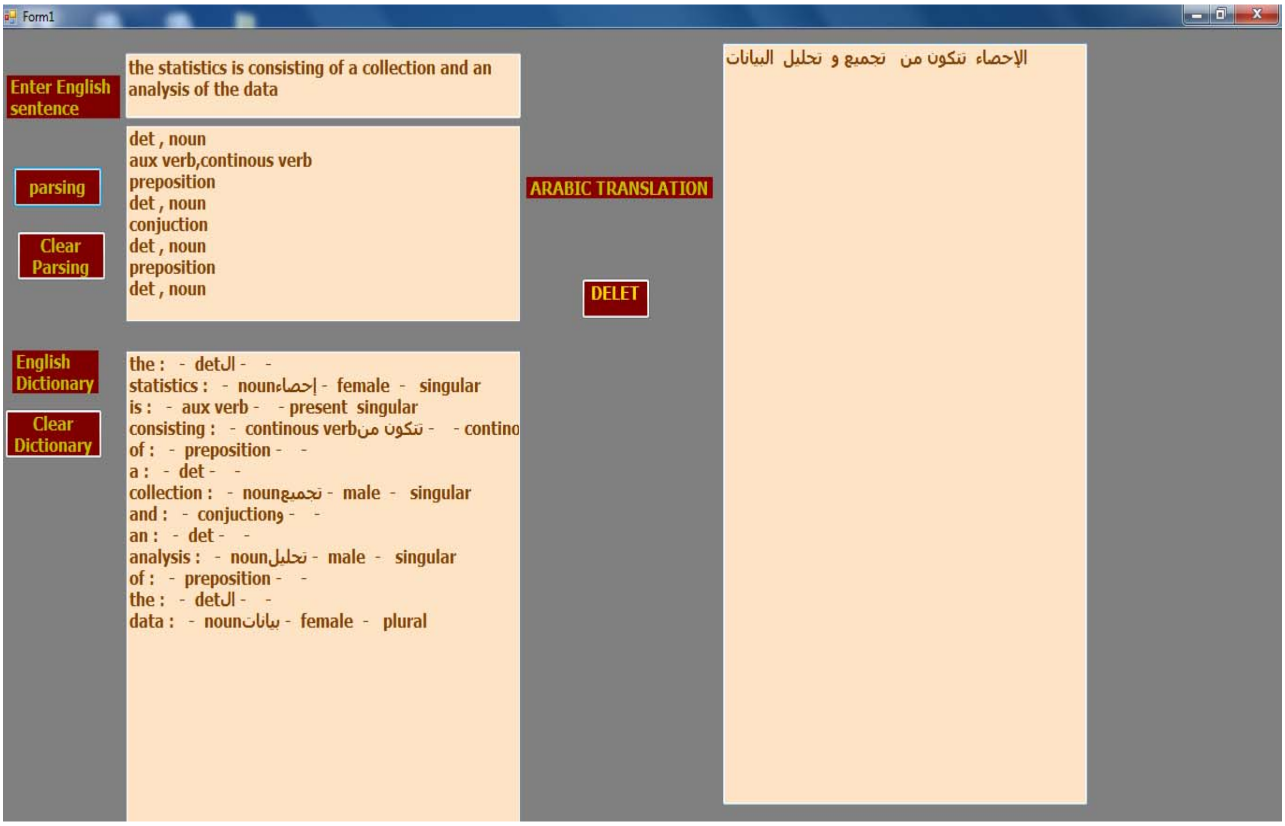
the mean is an average of the things

```
det , noun
verb to be , det , noun
preposition
det , noun
```

```
the :  -  detال -   -
mean :  -  nounوسط -  male  -  singular
mean :  -  nounوسط الحسابي -  male  -  singular
is :  -  verb to beيكون -   - present  singular
an :  -  det -   -
average :  -  nounمتوسط -  male  -  singular
of :  -  preposition -   -
the :  -  detال -   -
things :  -  nounأشياء -  male  -  plural
```

الوسط الحسابي يكون متوسط الأشياء

(b)



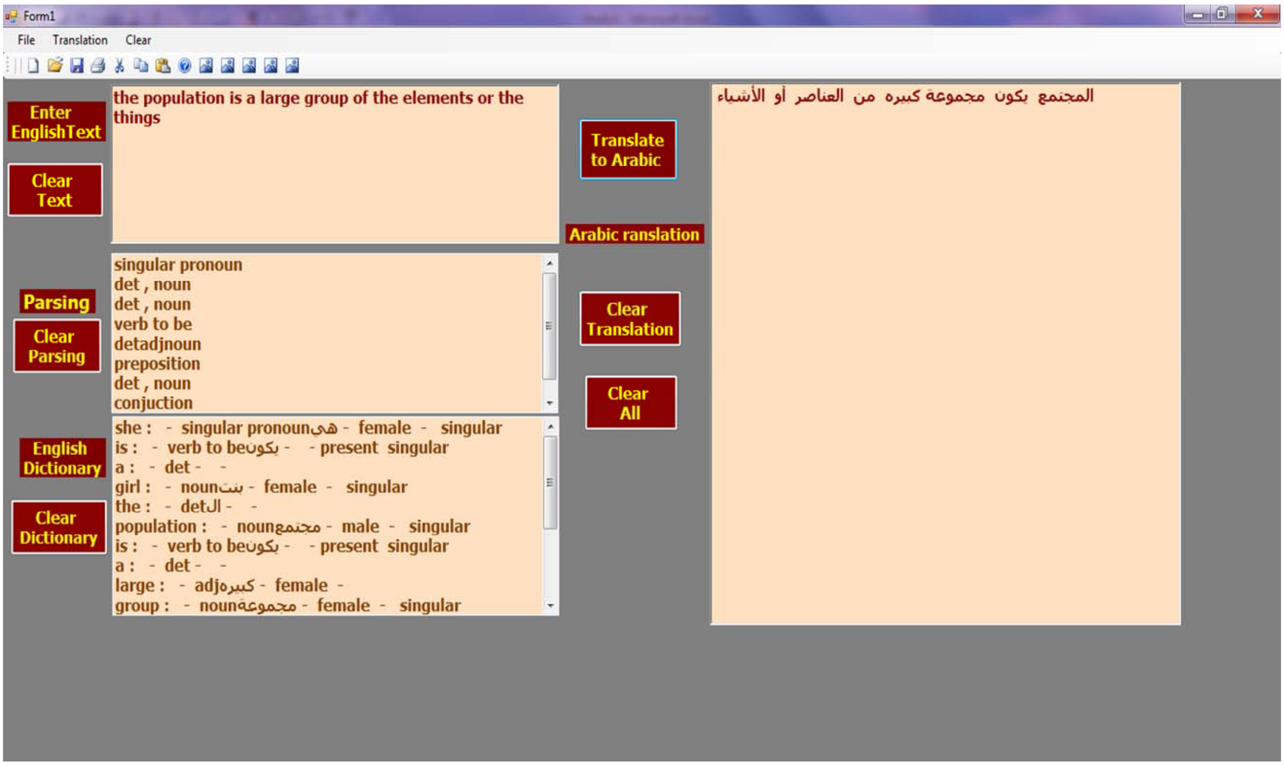
the statistics is consisting of a collection and an analysis of the data

```
det , noun
aux verb,continous verb
preposition
det , noun
conjuction
det , noun
preposition
det , noun
```

```
the :  -  detال -   -
statistics :  -  nounإحصاء -  female  -  singular
is :  -  aux verb -   - present  singular
consisting :  -  continous verbتتكون من -   - contino
of :  -  preposition -   -
a :  -  det -   -
collection :  -  nounتجميع -  male  -  singular
and :  -  conjuctionو -   -
an :  -  det -   -
analysis :  -  nounتحليل -  male  -  singular
of :  -  preposition -   -
the :  -  detال -   -
data :  -  nounبيانات -  female  -  plural
```

الإحصاء تتكون من تجميع و تحليل البيانات

(c)



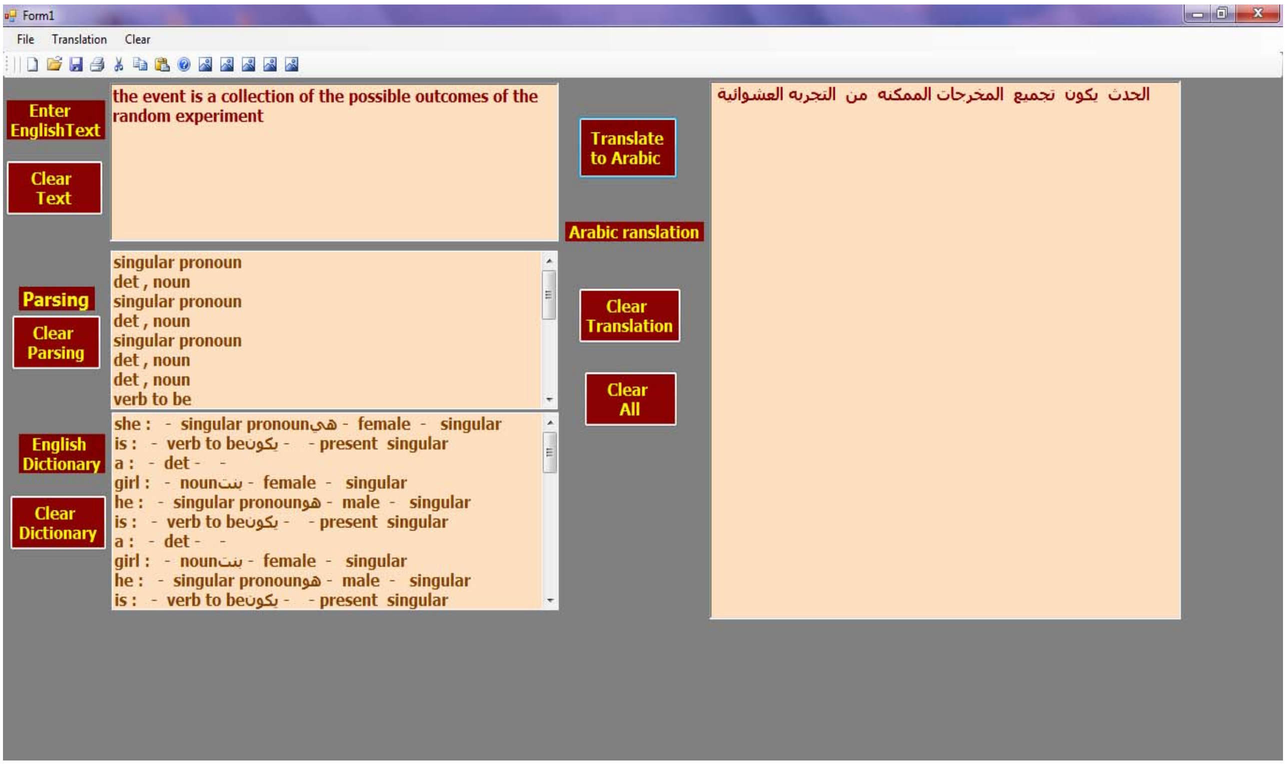
the population is a large group of the elements or the things

```
singular pronoun
det , noun
det , noun
verb to be
detadjnoun
preposition
det , noun
conjuction
```

```
she :  -  singular pronounهى -  female  -   singular
is :  -  verb to beيكون -    - present  singular
a :  -  det -    -
girl :  -  nounبنت -  female  -   singular
the :  -  detال -    -
population :  -  nounمجتمع -  male  -   singular
is :  -  verb to beيكون -    - present  singular
a :  -  det -    -
large :  -  adjكبيره -  female  -
group :  -  nounمجموعة -  female  -   singular
```

المجتمع يكون مجموعة كبيره من العناصر أو الأشياء

(d)



the event is a collection of the possible outcomes of the random experiment

```
singular pronoun
det , noun
singular pronoun
det , noun
singular pronoun
det , noun
det , noun
verb to be
```

```
she :  -  singular pronounهى -  female  -   singular
is :  -  verb to beيكون -    - present  singular
a :  -  det -    -
girl :  -  nounبنت -  female  -   singular
he :  -  singular pronounهو -  male  -   singular
is :  -  verb to beيكون -    - present  singular
a :  -  det -    -
girl :  -  nounبنت -  female  -   singular
he :  -  singular pronounهو -  male  -   singular
is :  -  verb to beيكون -    - present  singular
```

الحدث يكون تجميع المخرجات الممكنه من التجربه العشوائية

(e)

Figure 8. (a) Enter the English sentence; (b) shows the translation of the sentence "the mean is an average of the things"; (c) shows the translation of the statement "the statistics is consisting of a collection and an analysis of the data"; (d) shows the translation of the statement "the population is a large group of the elements or the things"; (e) shows the translation of the statement "the event is a collection of the possible outcomes of the random experiment".

6. Conclusion

At that work authors propose an approach for English-to-Arabic translation based on NLP which fills the gap in the field of scientific translation. The transformer approach employed in the proposed system combines some English grammar rules, structure transformational rules and Arabic morphological synthesis rules. The system analyzed, designed and implemented using c#.Net and SQL server. The results are successful Arabic translations from scientific English sentences. The proposed system is characterized by its translations that have high syntactic and semantic quality. In addition, its simplicity and modularity, enables future modification and extendibility with ease. In the future work authors intend to make that prototype system more realistic.

REFERENCES

[1] T. Ahmad AL-Tanni and Eyad M. Hailt, "A Direct English-Arabic Machine Translation System," *Information Technology Journal*, Vol. 4, No. 3, 2005, pp. 256-261. doi:10.3923/itj.2005.256.261

[2] Azza Abd El-Moniem Mohamed, "Machine Translation of Noun Phrases from English to Arabic," Faculty of Engineering, Cairo University, Giza, 2000.

[3] F. Al-Anzi, K. Al-Zame, M. husian and H. AL-Mutairi, "Automatic English/Arabic HTML Home Page Translation Tool," King Soud University, Riyadh, 1997.

[4] D. Arnold, L. Balkan, S. Meijer, R. Lee and H. L. Sadler, "Machine Translation," NCC Blackwell Lt, London, 1994.

[5] H. M. O. Mokhtar, N. M. Darwish and A. A. Rafea, "An Automatic System for English-Arabic Translation of Scientific Text (SEATS)," Master of Science Thesis, Computer Engineering Department, Faculty of Engineering, Cairo University, Cairo, 2000.

[6] K. Shaalan, A. Rafea, A. A. Moneim and H. Baraka, "Machine Translation of English Noun Phrases into Arabic," *International Journal of Computer Processing of Oriental Languages*, Vol. 17, No. 2, 2004. doi:10.1142/S021942790400105X

[7] K. R. Beesley, "Arabic Finite-State Morphological Analysis and Generation," *COLING*-96 *Proceedings*, Vol. 1, pp. 89-94.

[8] M. Palmer, O. Rambow and A. Nasr, "Rapid Prototyping of Domain-Specific Machine Translation Systems," Springer-Verlag, Berlin, Vol. 1529, 1998, pp. 95-102.

[9] M. Raji Zughoul and A. Miz'il Abu-Alshaar, "English Arabic/English Machine Translation: A Historical Perspective," *Translators*' *Journal*, Vol. 50, No. 3, 2005, pp. 1022-1041.

[10] Natural Language Processing & Applications Syntax, Coxhead, 2007.

[11] J. S. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," 2nd Edition, Prentice Hall, New Jersey, 2007.

[12] http://www.atasoft.com/products/mutarjim_v2.htm.