# Building of Database for English-Azerbaijani Machine Translation Expert System

**Guliyeva Zarifa** guliyeva_z_y@hotmail.com
*Institute of Information Technologies*
*Azerbaijan National Academy of Sciences*
*Baku, AZ 1141, Azerbaijan*

**Abstract**

In the article the results of development of machine translation expert system are presented. The approach of translation correspondences defining is suggested as a background for creation of data base and knowledge base of the system. Methods of transformation rule compiling applied for linguistic knowledge base of the expert system are based on the defining of translation correspondences between Azerbaijani and English languages.

**Keywords**: Machine translation expert system, data base, knowledge base, Azerbaijani.

## 1. INTRODUCTION

In this paper we present the research to be conducted for building first the bilingual data base then knowledge base of machine translation expert system from English to Azerbaijani. The expert system (ES) of machine translation based on morphological and syntactic knowledge has been created for practical application of the automatic dictionary in the machine translation system, and also checks acknowledgement of theoretically developed principles.

Developed system is realized on the basis of program Delphi 7 applied to the creation of control systems by databases and knowledge. The dictionary volume in the realized version makes 10 000 inputs on each language in the dictionary of the combined type. On the basis of the dictionary the database created for a concrete subject domain, however, at an initial stage is realized words of neutral lexicon for check of the rules which are a part of the knowledge base have been included. ES uses certain initial lexicon, grammar, and also semantics for creation interlinguistic models - interlingva representation (IR) of various word-combinations within the limits of a simple sentence. The given system concerns a wide spectrum of the information systems dealing with processing of texts in natural languages, in particular, in dialogue systems, providing dialogue with databases and knowledge bases in rather free natural language.

## 2. BRIEF DESCRIPTION Of AZERBAIJANI

Azerbaijani is a member of the Turkic branch of the Altaic language family. Specifically, it belongs to the Oghuz Seljuk sub-group (Akiner 1986), along with (Osmanli) Turkish and some dialects of Crimean Tatar (Campbell 1991). Other well known members of the Turkic branch include: Uzbek, Kipchak, Kyrgyz, Tatar, and Kazakh. The Turkic languages closely resemble each other and form a complex of mutually intelligible dialects.

Like all of the Turkic languages, Azerbaijani is agglutinative, that is, grammatical functions are indicated by adding various suffixes to fixed stems. Separate suffixes on nouns indicate both gender and number, but there is no grammatical gender. There are six nominal cases: nominative, genitive, dative, accusative, locative, and ablative; number is marked by a plural suffix. Verbs have voice, mood, tense, and nonfinite forms and they agree with their subjects in case and number, and, as in nouns, separate identifiable suffixes perform these functions. Subject-Object-Verb(SOV) word order in Azerbaijani is the norm, but other orders are possible under certain discourse situations. As a SOV language where objects precede the verb, Azerbaijani has postpositions rather than prepositions, and relative clauses that precede the verb. Azerbaijani has nine vowels and twenty three consonants. It also has Turkic vowel harmony in which the vowels of suffixes must harmonize with the vowels of noun and verb stems; thus, for example, if the stem has a round vowel then the vowel of the suffix must be round, and so on. In the research both linguistic systems of English and Azerbaijani were studied in details in order to establish full list of universals and differences between them for further determination of translation correspondences within all levels of the language.

## 3.MODEL OF TRANSLATION CORRESPONDENCES

As it is known basic notions of structural linguistics are mostly used in computational approaches. Analyzing the language for building computational model of a certain language one of the essential tasks is to define the set of feature and values relevant to its description. For processing of English-Azerbaijani texts in MT Expert system first there were defined properties of both languages, applying comparative analysis in order to define translation correspondences. After all morphological, morphonological and syntactic domains have been brushed through appropriate feature structures represents a whole set of basic structures.
After long lasting overview of existing models that can be used as a background of a practically new approach to the knowledge and data bases creation we suggested translation correspondences model which has certain advantages among others.

Translation as a specific process of inter language transformations concerns various language levels such as morphology, syntax, semantics and lexicology. In the translation complex interaction of these levels take place and new translation units such as translation correspondences appear as its result. These units refer to different language levels and this model reflects hierarchy of mentioned levels. Novelty of this model is that it can be placed in the centre of the whole model and modeling process. In the process of this model construction it was substantiated that most effective method to be taken as a guidance for building of both bases is method of correspondences selection. Elaboration of such method enables experts to choose most precise translation correspondences at all language levels and to provide optimum structure of expert system. While defining the notion of translation correspondences we proceed not only for translation equivalent from one language into the other because this factor refers not only to the lexical meaning of the word in the linguistic hierarchy. In the suggested approach determining of correspondences for grammatical categories, syntactic constructions and morpho-syntactic functions of words, word combinations and sentences is necessary condition providing adequacy and accuracy of translation. The wider spectrum of defined translation correspondences is with account of polysemy and multifunctional nature of language units the more complete the tokenized information filled into the data and knowledge bases would be, and thus it predetermines conditions for acquiring more accurate translation.
Translation process from one language into another is reduced to overcoming of divergences between languages. In MTES (machine-translation expert system) interlanguage divergences partially are taking away on each of analysis stages, and basically at a transfer stage. Difficulty of divergence overcoming is caused by that it is difficult to find translation correspondences between significant elements of languages with various structure.

## 4. INTERACTION OF CONSTITUENTS OF MT EXPERT SYSTEM

As it is known databases (DB) are most widespread technology for gathering, storage and processing of the huge data objects. However the latter do not allow to structure the data stored in them on the basis of the relations which exist between the facts directly in the real environment. Expert systems being large achievement of modern computer facilities and artificial intellect methods represent the specialized computer system capable to accumulation and generalization of experience of highly skilled experts. They also model reasoning of the experts in some certain area, using the knowledge base (KB)for this purpose, containing the facts and rules from this area and some procedure of a logic conclusion.
Industrially operating systems of machine translation give low quality translation and consequently require post editing. Experimental machine translation systems give more qualitative translation that essentially reduces a share of participation of the person in translation process. So, developed expert support system of machine translation works in an experimental mode and in this connection it can be characterized as experimental bilingual system of MT which uses full morphology, the limited syntax and partial semantics of applied languages. At the given stage of realization it is characterized by following properties:
• Full integrality of the descriptions of source and target languages. The principle of description integrality means that the morphology, syntax and the dictionary are completely co-ordinated with each other by the type of linguistic information replaced in it, and that this linguistic information in all three components is registered absolutely uniformly that is in the same formal languages.
• Declarativeness of the of linguistic knowledge set that is their total independence from algorithm. Declarativeness of the linguistic information set has two advantages at a stage of experimental operation of the system. Declaratively set linguistic model is easy to correct during machine experiments provided that the system simultaneously with translation of each phrase gives out a detailed protocol of its reception. This element is partially written down in expert support system of machine translation.

• Standardized nature of working language description formats. Both of working languages English and Azerbaijani are described under uniform schemes.
• Directivity of the lingware on one subject domain. This property of the system is a direct consequence of incompleteness of linguistic models of working languages. Morphology and syntax of working languages being at the initial stage of design are intended for processing of various scientific and technical texts, that is the wide spectrum of forms and constructions are considered to be met.
• Filling of dictionary database without directivity on a certain subject domain takes place, as for check of rules as a part of the system, word forms of the general and neutral lexicon were chaotically filled in database. Considering the fact that that the entries included in the dictionary are filled enough and fundamental in other words are capable to provide the adequate analysis and translation of the given word. The higher translation quality is the more full and more basic is the model of language composing a linguistic component of machine translation system.
Two constituents of expert system such as database and knowledge base are presented in appropriate ways and their brief description is suggested below.

4.1 **Database of MT expert system**
Database incorporates automatic dictionary as a part of expert system which represents the storehouse of the tokenized information used for text processing. Application and use of this information is possible only on the basis of its interaction with knowledge presented in knowledge base in the form of transformation rules of recognition and generation for grammatical, phonetic and semantic language phenomena. For compiling of automatic dictionary as foundation of database of the machine translation expert system the following principles were proposed below:
1. Comparison of genealogical origin of working pair of languages.
2. Typological comparison and identification if universals and differences of the languages.
3.Determining of lexical staff of the dictionary
4. Lemmatization(selection of lemmas) and glossary compiling
5. Selection of formal attributes of morpho-syntactic systems of both languages to define translation
correspondences
6. Formation of dictionary units for each part of speech.

The automatic binary dictionary is developed as a part of the integrated translation system and used for performance of the following problems:
- Serves as the basic tool of search (establishment) of lexical translation equivalents in ES;
- For work in a dialogue mode the dictionary is integrated into the general lexicographic base of ES and is the main informatively-directory base;
In ES as in one of systems of automatic text processing, automatic dictionary is a source of the grammatical information necessary for work of algorithms of automatic morphological and syntactic analyses, and also for work of lemmatization algorithms and knowledge base rules. The latter ensure functioning of the dictionary at the performance of the named functions in any paradigmatic form of a word.

**4.2 Classification of rules in linguistic knowledge base.**
Knowledge base of expert system is presented by a set productive rules each of which consists of: antecedent (conditions) and consequent (result) [2]. On a simple language of the user the rule consists of the right and left part. Knowledge represents a complex corrected unification (tree-based) grammar which includes in the structure elements different grammars, such as: context-free grammar (CFG) which is providing the morphological analysis and synthesis and being a basis of analyzer, linear grammar (LG) and constituent grammar which is providing morphological parse and synthesis. So, elements of CFG formalize the description of language model as formal grammar with finite-state set. Elements of LG fix a sequence of chain objects of formal-language model, that is the linear sentence structures of formal language model set in terms of grammatical classes of words. In the system "left to right" analysis strategy is applied: search of words, check of conditions, presence or absence of changes on conditions and addition of missing elements formally represent computer realization of finite-state grammar or CFG constructed on LG.
Formally transformation rules constitute foundation of knowledge base of expert system, and provide its functioning. Knowledge base can be divided into 3 blocks which in turn continuously co-operate with a database as a part of the analyzer of expert system. These blocks are possible be classified on:
- Recognition rules is a block of rules where sentences in a source language passes after identification of availability of word forms or word-combinations in the automatic dictionary. There by means of recognition rules an establishment of grammatical forms of words and branch of

suffixes or other grammatical indicators take place. For example, it is possible to carry the description of plural nouns formation to recognition rules in the English language, coming to light by means of detection of those or other affix changes, at performance of the set conditions(Figure 1)

- Substitution rules block. The found translation equivalents are replaced by substitution rules into syntactic chained structures. The way of representation of a syntactic sentence structure at which groupings of the words connected with each other are allocated, is called its system of components. Substitution rules are written down on the basis of combinability variations of parts of speech, and in an operating time of the syntactic block the established grammatical forms of sentence components pass to the stage of syntactic processing.(Figure 2 and Table 1)
- Generation rules at last stage of text processing carry out synthesis of sentence components in target language. In this block rules there is the construction of chains each part of which is appropriate part of speech with determined grammatical attributes. All lexico-grammatical processing procedures which took place in previous blocks are finally synthesized altogether and display the result generated in target Language.(figure 3)

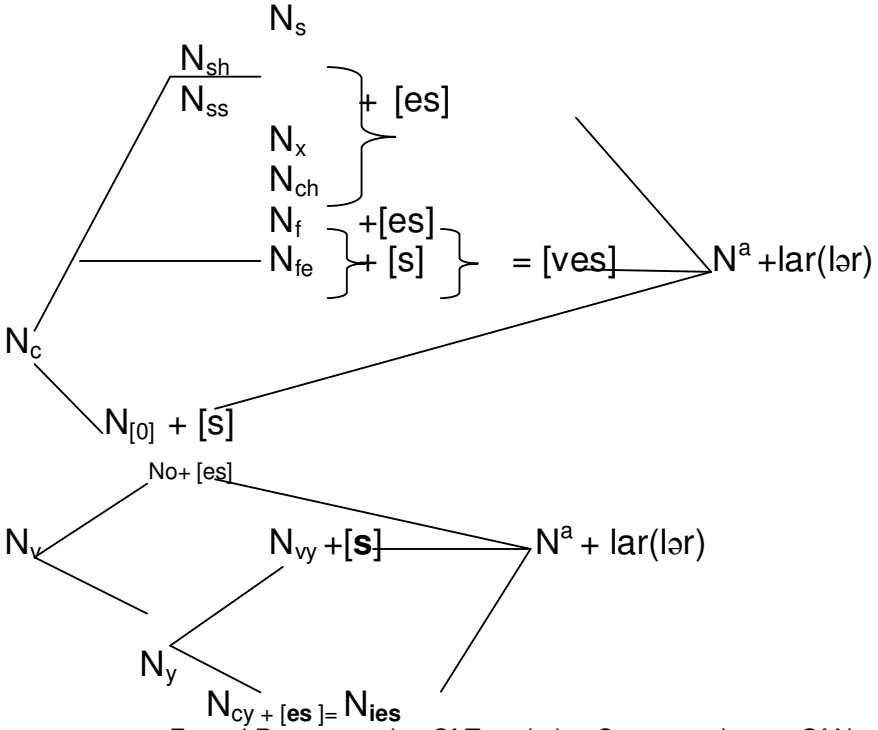


**FIGURE 1**. Formal Representation Of Translation Correspondences Of Nouns In Plural In Translation From English Into Azrebaijani.

Figure 1 schematically demonstrates the translation from English into Azerbaijani. So, let us open the given coding:

A noun ends with consonants/vowel phoneme and takes affix in the following cases:

- if noun ends with -s, - ss, -sh, -ch, -x then it takes affix [-es];
- if noun ends with -f it takes affix [-es] and word form ends [ves];
- if noun ends with -fe it takes affix [-s] in consequence of what the word form ends with [ves];
- if noun has zero flexion it takes affix [-s];
- if noun ends with –o it takes affix [-es] (the list of exceptions is included in database);
- if noun ends with –y it takes affix [-s] in combination of -y with vowel, but in combination of –y with consonant, the word form takes affix [-es] in consequence of which the word form ends with [-ies]

With provision for all above named conditions required for plural word form recognition translation equivalent of English word in a target language is found in the database for generation of corresponding word forms where the affix -**lar** is added to Azerbaijani translation equivalent, that is real for for words both with consonant and with vowel.

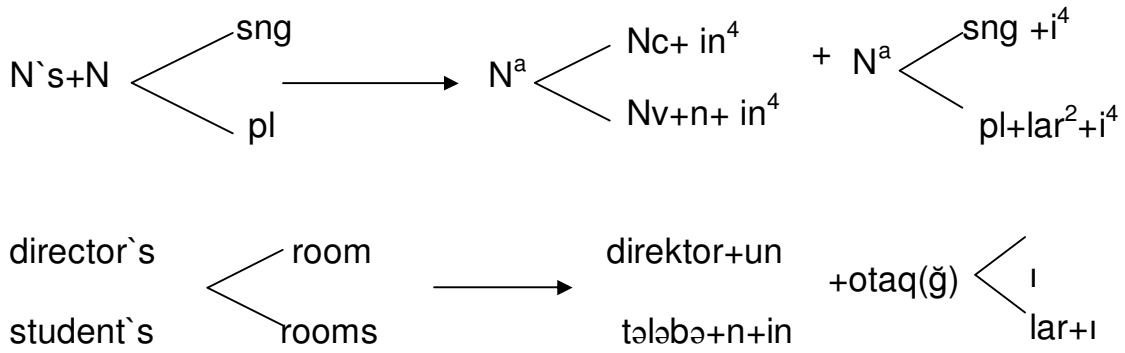


**FIGURE 2**. Translation Correspondences of Word Combination of Plural And Singular Nouns In Common And Possessive Cases

In figure 2 the combinability of possessive noun with plural and singular nouns is shown as Azerbaijani is agglutinative language and the connection between words in the sentence is expressed by affixes. In the left part of the fraction we see that Azerbaijani noun $N^A$ is also divided into $N^C$ (noun with cnonsonant in the end) and $N^V$ (with final vowel). This kind of distinguishing is important point in Azerbaijani grammar.

| **English verb tense formal representation** | | **Azerbaijani translation correspondences of English verb tense** |
|---|---|---|
| Present Simple<br>Present Continuous | S + $V_I$ /$V_s$<br>S+ (am, is, are)+ Participle I | S+(imperative)$V^A$+**ır**$^4$**(yir)**$^4$ +personal affix |
| Present Perfect | S+ have + Participle II | S+(imperative)$V^A$+**dı**$^4$ +personal Affix |
| Past Simple | S+ $V_{II}$/$V_{ed}$ | S+(imperative)$V^A$+**ib**$^4$**(mış)**$^4$ +personal Affix |

**TABLE 1**. Azerbaijani Translation Correspondences of English Verb Tense Formal Representation

Apart from above mentioned categories of English verb tenses were matched with their correspondences in Azerbaijani. Some of them found their translation correspondences but those which were not matched to the grammatical forms which can be substituted in output language. In the AD English verb is given in the form of infinitive without particle ***to*** but in Azerbaijani entry it is given as imperative form of the verb($V^a$) without its infinitive affix **–maq, -mək.** This way of verb presentation facilitates further synthesis of output text.

Translation correspondences for combinations of noun with prepositions which are common in English are expressed by 6 case affixes in Azerbaijani and it causes great difficulty for getting adequate translation. So, we matched prepositions to the 6 cases available in output language. This process is also important for data of prepositional verb constructions because English texts are full of such combinations and their recognition would be easier if prepositional verbs would be coded as united word form by its possession to appropriate POS and also coded by its requirement of a certain case affix depending on the meaning it carries.[11]

| **Azerbaijani case** | **English preposition** |
|---|---|
| Genitive case | Across, before, into, of, off, around, past, under |
| Dative case | To, for, towards |
| Accusative case | By |
| Instrumental (ablative) Case | Against, at, in, inside, on, during |
| Prepositional Case | About, after, except, from, out of, since, through, within |

**TABLE 2.** Translation correspondences of English prepositions to the noun cases in Azerbaijani

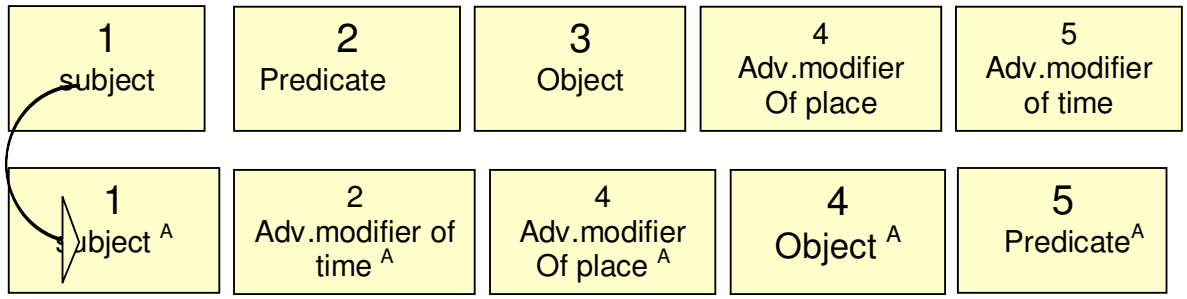


**FIGURE 3.** Difference in Location of Parts of Sentences in English and Azerbaijani Simple Sentences.

As it is seen the table 1 in presents a few verb tenses we have identified that are corresponding for English verb tenses. Modal verb and different verb combinations are expressed in Azerbaijani by means of analytical forms of verbs.
Figure 3 illustrates difference between word order of English and Azerbaijani(marked with $^{A}$). As it was mentioned in paragraph 2 Subject-Object - Predicate word order as in any Turkic language emerge many difficulties in tokenization simple sentences with participle I and II and their constructions, Gerundial and Infinitive constructions, as there is no similar category of gerund in Azerbaijani and translation correspondence of gerund is expressed in various ways, according to the construction.

## 5. CONCLUSION AND FUTURE WORK

As automatic text processing systems such as machine translation expert systems are advanced systems and include capabilities such as data conversion, finite-state transducer incorporating, collecting and processing numerous character sets, storage of huge volume and information, text retrieval and etc., English- Azerbaijani MT Expert System is of great demand hitherto, and as Azerbaijani can be referred to the list of lesser-studied languages many problems concerning formalization of this language are still remain unsolved.
So, MT expert system is directs to provide adequate translation scientific and technical texts from English into Azerbaijani. Recently the analysis of scientific texts from both language sources have been conducted in order to determine most frequently used grammatical and lexical constructions. In this foreshortened we are planning to extend the database of the system to 40 000 word forms and combinations and compile and input additional number of rules into the knowledge base for maintain more correct and accurate translation.

## 6. REFERENCES

[1] Makhmudov M.A. *"System of automatic text development*". (in Azerbaijani)– Bakı, 2002. 242 p.

[2] Veliyeva K.A. *"Avtomatic analysis and synthesis of the text* "(in Azerbaijani) – ADD, B., 1996.156 p.

[3] Veliyeva K.A. *"Problem of machine translation in turkology."*(in Azerbaijani) – *Researches*, 1, B., 2002.

[4] Marchuk Y.N. "*Translation modeling methods*" M.,Nauka,1985.(in Russian)

[5] Makhmudov M.A. *"System of turkish text automatic processing on lexico-morphological level"* B., 1991.(in Azerbaijani)

[6] Melchuk I.A. *"Theory experience of «Sense <-> Text » linguistic model."* – M.: Nauka, 1974.(in Russian)

[7] Bruderer H. *"The present state of Machine and Machine-Assisted Translation."* – OLB. V.1.

[8] Kemal Oflazer. "*Developing a morphological for Turkish*// Proc. of the NATO ASI on Language Engineering for
Lesser-studied languages. – NATO, ASI, JULY 2001, Ankara.

[9] Yilmaz *Kilichaslan*, Yuzlem Uchar. *A morpho-syntactic analyzer for Turkish sentences.* International XII Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN 2003.

[10] Shekal M., Gungur T., Kardesh O. *An approach for Machine Translation between Turkish and Spanish.* International XII Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN 2003.

[11] Guliyeva.Z.Y, Manafli.M. *Machine Translation Expert System for Texts in English-Azerbaijani Bilingual Environment.* International Conference on Artificial Intelligence, INISTA-2010, Turkey.