

Overcoming Agglutination Difficulties in the Development of an MT system from the Azerbaijani Language

Rauf Fatullayev,* Ali Abbasov,** and Abulfat Fatullayev***

*National E-Governance Network Initiative Project, Baku, Azerbaijan

**National Academy of Science, Baku, Azerbaijan

***Institute of Linguistics of National Academy of Science, Baku, Azerbaijan

fatullayev@gmail.com

ali@dcacs.ab.az

fabo@box.az

ABSTRACT

The paper presents the process of extraction of so called “active suffix chains” in the Azerbaijani language. The notion of an active chain is defined. A set of active chains for Azerbaijani is determined on the basis of statistical analysis of a large text corpus. The use of active chains for morphological analysis of Azerbaijani as well as for machine translation is presented.

1. MT approaches and the Azerbaijani language

Being one of the largest language groups the Turkic language group consists of about 60 languages (modern Turkish, Azerbaijani, Kazakh, Uzbek, Turkmen, Tatar, Kyrgyz etc.). About 40 of these languages are living, but 15 of them are dead languages [1].

However, majority of these languages (except modern Turkish) are less investigated. NLP (Natural Language Processing) systems (spell-checkers, speech recognition, understanding and generation, machine translation, information extraction, information retrieval, automatic summarization, spoken dialogue systems etc.) resources and tools are very rare for these languages.

Development of machine translation (MT) systems is one of the foreground directions of NLP. From the point of view of the development of the MT systems (including other NLP systems) Turkish is in better condition than other Turkic languages [2–8]. Some MT systems: ProÇeviri (www.proceviri.com), TranSphere (www.apptek.com), Babylon (www.babylon.com), SameTran (www.sametran.com) have the translation module into Turkish but none of them has the translation module from/into any other Turkic language. Although some research has been carried out in the field of MT from/into Turkish dedicated to different aspects of the development of the MT system for Turkish the whole technology for the creation of an MT system from Turkish has not also been created yet.

Existing approaches to the development of the machine translation systems can be divided into three groups: rule based MT, statistic based MT and example based MT

[9–11]. Apart from these approaches, there are also hybrid methods which use a combination of the aforesaid approaches and other heuristic methods in order to increase the quality of translation [12, 13].

Statistical MT is now the mainstream approach that requires appropriate resources (e.g. bilingual corpora) for the development of an MT system [14, 15]; however it is a hard task to create the bilingual corpora for so-called “low density” languages – that is, languages where few on-line resources exist.

Most Turkic languages (including Azerbaijani) are low-density languages and the bilingual corpora for these languages have not been created yet. Due to this fact Google’s translation system doesn’t include any Turkic language. The absence of a strict word order and the fact that one word may be translated as a whole sentence makes it even more difficult to use the statistic-based and example-based approaches while creating MT systems for the Turkic languages [16, 17].

For these reasons we came to the conclusion that at present the rule-based approach is more “attractive” for the creation of an MT system for Azerbaijani and we choose the rule based MT approach for the creation of the MT system for Azerbaijani.

While developing MT systems from the languages of the Turkic group on the basis of the rule-based MT approach some problems related to the agglutinative nature of these languages arise. In Turkic languages, every word-form morphologically consists of the stem and the simple or compound suffixes. Languages of the Turkic group are morphologically rich and are characterized by highly productive morphological processes that may generate a very large number of word forms for a given stem. Modeling each word-form as a separate lexical unit leads to a number of problems for the development of formal linguistic technologies such as machine translation, speech recognition, text to speech etc. systems.

Despite the vocabulary differences, the morphological structures (e.g. similar rules of the word formation, existence of the simple suffixes with the similar function) and syntactic structures (e.g. the similar rules for formation of the noun and verb phrases, the same word order) of the languages of the Turkic group are very close. Therefore it is easier to develop machine translation systems among the languages of the Turkic group than those from/into the languages not belonging to this group.

Since the grammatical structures of these languages are very close, it is possible to develop a word-to-word MT system that uses simple multilingual dictionaries and the database of equivalent simple suffixes [18–20]. Examples of such structures are shown in Tables 1 and 2 respectively.

For example, the below sentences are equivalents of the English sentence “My little son goes to school” in six Turkic languages. Let us notice that the correct translation between those sentences in most cases merely consists in the appropriate replacement of word stems and suffixes according to Tables 1 and 2.

| | |
|--|------------|
| <i>Mən-im kiçik oğl-um məktəb-ə ged-ir</i> | (Azerb.), |
| <i>Ben-im küçük oğl-um okul-a gid-iyor</i> | (Turkish), |
| <i>Men-inq kiçik ўғл-ум мактаб-қа бора-ди</i> | (Uzbek), |
| <i>Men-in bəlekey bala-m mektep-ke bara jat-ır</i> | (Kazakh), |
| <i>Men-in kiçine uul-um mektep-ke bar-dı</i> | (Kyrgyz), |
| <i>Min-em keçkene ul-um məktəp-qə bar-a</i> | (Tatar). |

Table 1. Turkic multilingual dictionary

| Azerb. | Turkish | Uzbek | Kazakh | Kyrgyz | Tatar |
|--------|---------|--------|---------|--------|---------|
| mən | ben | men | men | men | min |
| kiçik | küçük | kiçik | bəlekey | kiçine | keçkene |
| oğl | oğl | ўғл | bala | uul | ul |
| məktəb | okul | maktab | mektep | mektep | məktəp |
| ged | gid | bor | bar | bar | bar |

Table 2. Database of equivalency of simple suffixes

| | | | | | |
|----|------|-----|----|----|----|
| im | im | ing | in | In | em |
| um | um | im | m | um | um |
| ə | a | qa | ke | ke | qe |
| ir | iyor | di | ır | dı | a |

However for the development of the MT system from Turkic languages into (for example) English (analytical language) or Russian (inflectional language) we have an essentially different situation.

In these cases we should translate suffix chains in whole (without separating them into simple suffixes) and this fact causes some problems. By adding various suffixes to the stem of the same verb, it is possible to create 17947 word-forms in Tatar [21], 11390 in Turkish and 13592 in Uzbek [22]. In the Kazakh language, the number of suffixes that create various word-forms from noun and verb stems is about 500 and 1000 respectively [23]. In Azerbaijani, the number of word-forms formed from the same stem is more than 8000 [24].

For these two reasons (a large number of suffix chains and the necessity to translate suffix chains in whole), we find it reasonable to define the subset of suffix chains used in texts instead of the set of all possible suffix chains. In other words we should take into account that the frequency with which all these suffixes and their chains occur is not the same.

If this fact is considered while creating an MT system, then the difficulties related to a large number of suffix chains can be avoided: not all suffix chains, but the suffix chains used in real texts can be determined and included in the database of suffix chains.

Suffix chains regularly found in texts are called active suffix chains and our purpose in this paper is to define the subset of active suffix chains for Azerbaijani.

Hereinafter Azerbaijani is taken as a source and English as a target language.

2. Suffix chains and their translations

In this section we will show the necessity for taking suffix chains in whole in the translation process.

We will call the number of the simple suffixes that comprise a suffix chain the *length* of this chain. Simple suffixes will also be referred as a suffix chain (whose length is one).

In the grammar of the modern Azerbaijani language, suffixes are divided into two groups – lexical and grammatical suffixes [8]. Lexical and grammatical suffixes form various word-forms from the same word stem by joining word stems in a certain sequence (for example, it is possible to form the word-forms *ev-də*, *ev-dəki-lər-in*, *ev-də-dir-lər*, *ev-dəki-lər-dən-siniz-mi* and etc. from the stem *ev*, Table 3).

Table 3. Database of the active word-forms (fragment)

| Word-form | Word-form | Word-form | Word-form | Word-form |
|-----------|------------|----------------|--------------|---------------|
| 1. Ev | 6. ev-idir | 11. ev-ində | 16. ev-inin | 21. ev-lərdən |
| 2. ev-dir | 7. ev-imdə | 12. ev-lərinə | 17. ev-inə | 22. ev-ləri |
| 3. ev-də | 8. ev-imin | 13. ev-indədir | 18. ev-lə | 23. ev-lərin |
| 4. ev-dən | 9. ev-imə | 14. ev-indən | 19. ev-lər | 24. ev-lərinə |
| 5. ev-i | 10. ev-in | 15. ev-inizə | 20. ev-lərdə | 25. ev-lərinə |
| ... | | | | |

A question arises: is it possible to get the correct translation of a word-form by translating each simple suffix of its suffix chain separately and putting these translations together? It is a very important question from the point of view of the development of the MT system from Azerbaijani, but as we will see below, the answer to this question is negative.

Let's have a look at the examples in Table 4 (suffix chains and their translations are underlined in the same way). While the suffix *-də* is translated as *at* and the suffix *-dir* is translated as *he/she/it* is in the first example, in the second example, the suffix *-dir* cannot be translated separately. Here the suffix chain *-dir-lər* should be translated as *they are*. In the third example, the suffix chain *-dir-lər* is translated as *they are* and the plural suffix *-s* which is added to the end of the word. In the fourth example, different translation takes place and the suffix chain *-dir-lər-mi* is translated as *are they* with the

Table 4. Examples of the translation of suffix chains into English

| Word-form | Stem of the word-form | Suffix chain | Translation of word-form |
|---------------------------|--------------------------|-------------------|--------------------------|
| 1. <i>evdədir</i> | <i>ev</i> home | <i>də-dir</i> | he is at home |
| 2. <i>evdədirilər</i> | <i>ev</i> home | <i>də-dir-lər</i> | they are at home |
| 3. <i>tələbədirilər</i> | <i>tələbə</i> student | <i>dir-lər</i> | they are students |
| 4. <i>tələbədirilərmi</i> | <i>tələbə</i> student | <i>dir-lər-mi</i> | are they students |

Table 5. Examples of the translation of suffix chains into Russian

| Azerbaijani suffix | Russian equivalent |
|---|---|
| <i>-ir, -ir, -ur, -ür</i> (<i>ged-ir</i> – he goes) | <i>-em, -ëm, -um</i> (<i>ид-ем</i>) |
| <i>-lär, -lar</i> (<i>kitab-lar</i> – books) | <i>-и, -ы, -а, -я</i> (<i>книг-и</i>) |
| <i>-ir-lär</i> (<i>ged-ir-lär</i> – they go) | <i>-ут, -ют, -ат, -ят</i> (<i>ид-ут</i>) |

plural suffix *-s* added to the end of the word-form. The number of these examples can be increased.

It is worthwhile to note, that similar phenomena take place not only for English but for other non-Turkic languages as well (Table 5).

As it can be seen from the examples, the compositional translation of each suffix included in the suffix chain can lead to erroneous results. In order to get the right translation, the suffix chain should be translated as a whole.

We introduce the notion of active suffix chains. Active suffix chains are suffix chains that occur regularly in real texts. Therefore for the development of an MT system, first we define the set of active suffix chains.

3. Active suffix chains of Azerbaijani

In this section we will define the set of active suffix chains for Azerbaijani.

Creation of the suffix chains databases is also necessary for the morphological parsing. One of the main purposes of the formal morphological analysis in Azerbaijani is to separate the stem of the word-form from its suffix chain. Because like in all agglutinative languages, grammatical relations among word-forms in the Azerbaijani sentence are determined by suffix chains. Correct determination of the suffix chain seriously impacts the correct course of the further analysis process.

We should especially point out here that when we say suffixes and suffix chains, we only mean grammatical suffixes and suffix chains formed from them (Number of simple grammatical suffixes of Azerbaijani is about 100 [25]). We are not examining simple lexical suffixes or lexical suffixes in suffix chains. Because word-forms formed by lexical suffixes are kept as separate lexical units in the dictionary of an MT system (this refers both to prefixes – *na, bi, ba, la, a, anti* and to other lexical suffixes *-li, -li, -lu, -lü, -çl, -çl, -ç, -çü, -lıq, -lik, -luq, -lük* etc.).

For example, although the words *balıq* (fish), *balıq-çl* (fisherman) and *balıq-çl-lıq* (fishery) come from the same stem, all three words are kept in dictionary as units. This takes place because formally there are no general rules to generate the translation of the word-forms *balıq-çl* (fisherman), *araba-çl* (wagoner) from the translation of the words *balıq, araba* (wagon).

Definition of the set of active suffix chains is carried out in two steps: The first step works on a small corpus, the other on a large corpus. First, we manually marked the

corpus of 300 000 word-forms (by separating stems of the word-forms from their suffix chains). Since we did not have any set of suffix chains at the beginning, first we decided to define most frequently used chains in the small corpus. Afterwards we automatically verified and improved the obtained results (the set of chains received in the first step) against the large corpus of 12 000 000 word-forms.

4. Definition of the initial set of active suffix chains

The number of word-forms in the small corpus was about 39000 (after some processing). These word-forms were put in the database and stems of these word-forms were separated from their suffix chains manually in the way presented in Table 3.

During this process suffix chains were encountered 111406 times, but the number of various suffix chains was 1415 (including variants of the suffix chains with the same meaning). These suffix chains form the basis of the “Database of suffix chains” of the Azerbaijani-English MT system. After grouping these chains (the ones that have the same function, but different spelling, for example, *acaq*, *acağ*, *əcək*, *əcəy*, *yacaq*, *yacağ*, *yəcək*, *yəcəy*), the number of suffix chains diminished to 627.

The length of suffix chains was also calculated during the computer analysis. The frequencies of the use of the suffix chains in the order by their length are given in Table 6. As we can see from these results, very long suffix chains are rare and such chains almost are not used in real texts. This can be clearly seen from Table 6. Suffix chains longer than five simple suffixes were not encountered in the analyzed texts. The following table shows the frequency with which suffix chains are used by their length.

One of the possible reasons why suffix chains longer than five simple suffixes are not encountered could be that we did not take into account the lexical suffixes. On the other hand, the fact that long suffix chains are not used indicates that although the use of such suffix chains is principally possible, they are not used in writing. If it is necessary, the idea to be expressed by means of a long suffix chain is expressed by a shorter suffix chain (or words) that have the same meaning. For example: the sentence “*Siz bizim dəvət et-di-k-lər-imiz-dən-siniz-mi*” (Are you one of the people who we invited?) is replaced with an equivalent sentence “*Biz Sizi dəvət et-miş-ik-mi*” (Have we invited you?) or another similar equivalent sentence, for example, with the sentence “*Siz dəvət edil-mi-siniz-mi*” (Have you been invited?). A chain that has seven simple suffixes is replaced with a chain that has three simple suffixes.

As it can be seen in Table 6, a chain of five simple suffixes was encountered five times (0.004% of all cases), a chain of four simple suffixes was encountered 223 times (0.200%), a chain of three simple suffixes was encountered 6,895 times (6.189%), a chain of two simple suffixes was encountered 41,331 times (37.099%) and a chain of one suffix was encountered 62,952 times (56.507%).

Table 6 also shows the distribution of suffix chains that have the same functions disregarding repetition.

The number of distinctive chains of five simple suffixes was four, the number of chains of four simple suffixes was 66, the number of chains of three simple suffixes was 248, and the number of chains of two simple suffixes was 257 while the number of chains of one simple suffix was 53.

Table 6. Frequency of suffix chains

| Length of chain | Frequency | Percentage of repeat | Unrepeated chains |
|-----------------|-----------|----------------------|-------------------|
| 5 | 5 | 0.004% | 4 |
| 4 | 223 | 0.200% | 66 |
| 3 | 6,895 | 6.189% | 248 |
| 2 | 41,330 | 37.099% | 256 |
| 1 | 62,952 | 56.507% | 53 |
| Total | 111,405 | 100.000% | 627 |

Basing on these figures, we can say that most of the suffix chains used in the Azerbaijani language consist of one, two or three simple suffixes.

These suffix chains comprise 99.795% of all most frequent suffix chains. Relative long suffix chains (the ones that have four, five simple suffixes and longer) compound only 0.205% of all chains.

5. Verification and improvement of the results

The completeness of the set of chains obtained in the previous section was verified again within the text corpus of 12 million word-forms by using morphological analyzer for Azerbaijani. This process is carried out in the following order.

First different word-forms of the corpus are defined and included in the word-form database. Further all word-forms have been run through the morphological filter (as in Example 1). The word-forms which are correctly separated into the stem and suffix chain are excluded from the database (Because stems and suffix chains of these word-forms are already included in the MT dictionary and suffix chains database correspondingly). The word-forms which have remained in the database are manually separated into the stem and suffix chain. Thus 196 new suffix chains are defined and the number of encountered suffixes reached 823.

Although the volume of the text corpus has increased 40 times, the number of suffixes has increased 1.31 times and the percentage of the use of longer suffix chains did not change. On the basis of this result it is possible to say confidently that active suffix chains of the Azerbaijani language do not exceed 1000.

Great volume of the analyzed text corpus allows us to say that the expansion of the text corpus will not cause a considerable change in relative frequency indicators.

Besides, the types of active suffix chains on the definition of their capability to join the stems belonging to the different parts of speech are determined basing on the parts of speech of words they are combined with. Along with well known ambiguity problems (lexical, syntactical etc.), agglutinative languages bear the grammatical ambiguity (ambiguity of suffixes) and this information can be used in the disambiguation process. 534 chains ($\approx 64.88\%$) of all chains can join only verb stems (verb chain), 254 chains ($\approx 30.86\%$) can join non-verb stems (non-verb chain) and 35 chains ($\approx 4.25\%$) can join both types of stems (dual chain). For the dual chains their frequency of occurrence as verb or non-verb chains is also defined and this statistics is also used in lexical and grammatical disambiguation process.

Table 7. Database of active suffix chains (Fragment)

| Suffix chain | Other variants of suffix chains | Structure of the chain | Type |
|---------------|---|------------------------|----------|
| <i>Am</i> | <i>əm, yam, yəm</i> | | |
| <i>da</i> | <i>də</i> | | <i>N</i> |
| ... | | | |
| <i>ir</i> | <i>ur, ür, yur, yür, ir, yir, yir</i> | | <i>V</i> |
| <i>uram</i> | <i>ürəm, yuram, yürəm, ıram, irəm, yıram, yirəm</i> | <i>ur-am</i> | <i>V</i> |
| <i>dadır</i> | <i>dədir</i> | <i>da-dır</i> | <i>N</i> |
| ... | | | |
| <i>lar</i> | <i>lər</i> | | <i>N</i> |
| <i>larda</i> | <i>lərdə</i> | <i>lar-da</i> | <i>N</i> |
| ... | | | |
| <i>mış</i> | <i>mış, muş, müş</i> | | |
| <i>mışdır</i> | <i>mışdır, muşdur, müşdür</i> | <i>mış-dır</i> | <i>V</i> |
| ... | | | |

So, we have defined the composition of the main information included in the database of the active suffix chains.

The fragment of this database is shown in Table 7.

The types of suffix chains are indicated in the 3rd column of Table 7. The letter “*V*” written in the third column shows that this suffix chain is a verb chain, while the letter “*N*” – non-verb chain. If none of these letters is written there, the suffix chain is a dual chain.

In the next section we consider the use of the database of the active suffix chains in the morphological analysis process of the Azerbaijani word-forms.

6. The use of the active suffix chains database in morphological analysis process

The formation of the grammatical relations among word-forms in a sentence may significantly differ for different languages. In analytical languages (for example: in English) the grammatical relations among word-forms in a sentence, in most cases, are defined by word order and/or prepositions. Separate words don't have grammatical information and such information can only be acquired from the strict word order. But in agglutinative languages grammatical relations among word-forms in a sentence are formed by the rich set of suffix chains. In order to define the grammatical relations among the word-forms it is necessary to separate stem from suffix chain of the word-form to determine the presence of the participation of the word-form in syntactic structures of the sentence.

In agglutinative languages, formal (by computer) morphological analysis can be carried out on the basis of dictionaries prepared beforehand. The dictionary of the Azerbaijani word stems is also developed in frame of the project and the fragment of this dictionary is indicated in Table 8.

Table 8. Dictionary of Dilmanc MT system (Fragment)

| Stem | Part of speech | English translation |
|-------------------|----------------|---------------------|
| <i>tərcümə et</i> | verb | translate |
| ... | | |
| <i>dilmanc</i> | noun | translator |
| ... | | |
| <i>yaz</i> | verb | write |
| <i>yaz</i> | noun | spring |
| <i>yazı</i> | noun | record |
| ... | | |
| <i>qur</i> | verb | construct |
| <i>quru</i> | verb | dry |
| <i>quru</i> | adverb | dry |
| <i>quru</i> | noun | land |
| ... | | |
| <i>qorx</i> | verb | play |
| <i>qorxu</i> | noun | fear |
| ... | | |
| <i>cədvəl</i> | noun | table |
| ... | | |

Disregarding the ambiguity problems, we will schematically describe the work of the morphological analyzer for Azerbaijani.

The morphological analysis algorithm of word-forms in Azerbaijani is shown in [24]. The morphological analysis process can be described shortly as follows:

1. The whole word-form is sought in the dictionary of stems (Table 8).
2. If the word-form is not found in the dictionary, its last letter is discarded and the truncated part of the word-form is sought in the dictionary again. This process is iterated until the word-form or its truncated part is found in the dictionary of stems. The discarded part of the word-form is sought in the database of suffix chains (Table 7).
3. If the discarded part of the word-form is a suffix chain and this chain can join the stem of this word-form (for example, if the stem is a verb, then the type of the suffix chain should be V – a verb chain), then the algorithm stops, otherwise it loops back step 2.
4. After the stem and suffix chain of the word-form are defined, the word-form is provided with the information included in the databases of stems and suffix chains for its stem and suffix chain.

Example 1. Let's consider the formal morphological analysis process of the word-form *məktəbdədir* (he/she/it is in the school). Starting from the whole word-form all its reminders are sequentially sought in the dictionary of stems (Table 8). Only after 6 character deletions the stem *məktəb* of the word-form is found in the dictionary.

Discarded part *-dədir* is also found in the database of suffix chains. So, process is stopped and we can write

$$m\acute{a}kt\acute{a}bd\acute{a}dir \Leftrightarrow m\acute{a}kt\acute{a}b-d\acute{a}dir.$$

The word-form and its beginnings (according to above mentioned algorithm) with the discarded parts are shown below:

1. *m\acute{a}kt\acute{a}bd\acute{a}dir*,
2. *m\acute{a}kt\acute{a}bd\acute{a}di* *r*,
3. *m\acute{a}kt\acute{a}bd\acute{a}d* *ir*,
4. *m\acute{a}kt\acute{a}bd\acute{a}* *dir*,
5. *m\acute{a}kt\acute{a}bd* *\acute{a}dir*,
6. *m\acute{a}kt\acute{a}b* *d\acute{a}dir* ▲

Example 2. Let's carry out a formal morphological analysis of the word-form *qorxuram* (I am afraid). According to the morphological analysis algorithm, after 4 character deletions – the word-form *qorxu* is sought and found in Table 8. But the discarded part of the word-form – *ram* is not a suffix chain (Table 7). Therefore the process continues and after 5 character deletions – the word-form *qorx* (verb stem) is sought and found in Table 8, the discarded part – *uram* of this word-form is found in Table 7. The process stops because such a suffix chain is found and the type of the stem (verb type) corresponds to the stem (verb).

Steps of this process are presented below:

1. *qorxuram*,
2. *qorxura* *m*,
3. *qorxur* *am*,
4. *qorxu* *ram*,
5. *qorx* *uram*.

Thus, after this process we get

$$qorxuram \Leftrightarrow qorx-uram \blacktriangle$$

Example 3. For the word-form *yazır* (*yaz-ır*, he/she/it writes)

1. *yazır*,
2. *yazı* *r*,
3. *yaz* *ır*.

in the 2nd step process does not stop, because *-r* is not a suffix chain (though *yazı* is found in Table 8). In the next step two variants of the stem *yaz* (verb and noun) are found in Table 8. Since the discarded part – *ır* is a verb chain (Table 7) the algorithm chooses the verb variant of the stem *yaz* ▲

7. Dilmanc MT system

Research works on the development of Speech and MT technologies for Azerbaijani are being led since 2003 [26–27]. Most of the necessary works (development of the

MT dictionaries, creation of the formal grammar for Azerbaijani, MT algorithms from/ into Azerbaijani, synthesizer and analyzer algorithms of the Azerbaijani sentences, definition of the threephone set for the ASR system etc.) for the development of these technologies are carried out from scratch.

The research work presented in this paper is one of such important steps on creation of the applied linguistic technologies for Azerbaijani (All researches are carried out within the joint projects “Development of the MT system for Azerbaijani”, “Development of the Speech Recognition system for Azerbaijani” of the Ministry of ICT of Azerbaijan and UNDP-Azerbaijan).

Dilmanc is the 1st MT system for Azerbaijani developed in the frame of this project. Presently Dilmanc MT system can translate in three directions – Azerbaijani-English, English-Azerbaijani and Turkish-Azerbaijani (www.dilmanc.az) [28]. The MT system gives good enough – intelligible translations especially for texts pre-edited for the machine translation. The pre-editing rules are also developed and can be downloaded from www.dilmanc.az.

Dilmanc MT system (Fig. 1) has the following characteristics on each direction (all these items have been developed for the first time):

Azerbaijani-English direction.

1. MT dictionary of Azerbaijani word stems (≈ 120000 lexical units including word phrases and terms);
2. Database of the active suffix chains (≈ 1000 active chains);
3. Database of the formalized rules for the lexical and syntactical disambiguation in Azerbaijani (≈ 1500 rules);
4. Database of translations of the active suffix chains of Azerbaijani (≈ 2300 rules¹);
5. Formalized rules of the “traditional” grammar of Azerbaijani for the definition of the noun and verb phrases;
6. Formal morphological analysis algorithms of Azerbaijani word-forms;
7. Formal syntactic analysis algorithms of the Azerbaijani sentences;
8. Algorithms for the synthesis of the English sentences.

English-Azerbaijani direction.

1. English-Azerbaijani MT dictionary (≈ 115000 lexical units including word phrases and terms);
2. Database of the formalized rules for the lexical and syntactical disambiguation (≈ 1400 rules);
3. Database of the formalized rules for the synthesis of Azerbaijani suffix chains (≈ 300 rules);
4. Database of the rules for the POS disambiguation in the English sentence (≈ 90 rules);
5. Database of the rules for delimitation of clauses in the English sentence (≈ 40 rules);
6. Algorithms for the formal syntactic analysis of the English sentences.
7. Algorithms for the synthesis of the Azerbaijani sentences.

¹Some suffix chains subject to the part of speech of the word stem can be translated by different rules and as a result number of translation rules more than number of active suffix chains.

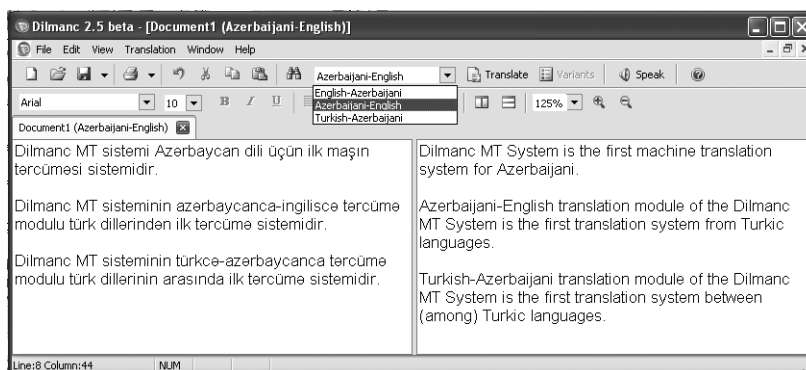


Figure 1. Dilmanc MT system.

Turkish-Azerbaijani direction.

1. Turkish-Azerbaijani MT dictionary (≈ 20000 lexical units);
2. Database of the equivalency of Turkish and Azerbaijani suffix chains (≈ 1000 chains).

8. Conclusion

The set of active suffix chains defined as the result of this research facilitate the translation process from Azerbaijani. Definition of this set yields opportunity to limit the set of Azerbaijani word-forms and thereby make it possible to create different applied linguistic technologies (including MT, ASR, TTS etc. systems) for Azerbaijani.

On the other hand the set of active suffix chains together with approach (explained in [23]) for the creation of the MT dictionary from Azerbaijani yields possibility to minimize the volume of the dictionary of the MT system and lead to the solution to the problem of the development of the machine dictionary from the Turkic languages (The Turkish-Azerbaijani direction of Dilmanc MT system can be a good example of this fact).

BIBLIOGRAPHY

- [1] A.A. Kibrik, E.R. Tenishev, E.A. Poceluevskij and I.V. Kormushin. 1997. Jazyki mira: Tjurkskie jazyki [Languages of the world: Turkic languages]. Moscow: Indrik. 542 p.
- [2] Cicekli I, Korkmaz T. 1998. Generation of Simple Turkish Sentences with Systemic-Functional Grammar. In: Proceedings of the 3rd International Conference on New Methods in Language Processing (NeMLaP-3), Sydney, Australia, January 1998.
- [3] Durgar-El-Kahlout I, Oflazer K. 2006. Initial Explorations in English to Turkish Statistical Machine Translation. Workshop on Statistical Machine Translation, New York, NY, June 2006.
- [4] Temizsoy M, Cicekli I. 1998. An Ontology-Based Approach to Parsing Turkish Sentences. In: Proceedings of AMTA'98-Conference of the Association for Machine Translation in the Americas, Lecture Notes in Computer Science 1529, Springer Verlag, October, Langhorne, PA, USA.
- [5] Tur G, Hakkani-Tur D, Oflazer K. 2000. Statistical Modeling of Turkish for Automatic Topic Segmentation. Bilkent University, Computer Engineering Technical Report BU-CE-0001, January.

- [6] Vural E, Erdogan H, Oflazer K, Yanikoglu B. 2005. An Online Handwriting Recognition System For Turkish. In: Proceedings of SPIE Vol. 5676 Electronic Imaging 2005, San Jose, January 2005.
- [7] Arisoy, E. Saraclar, M. 2007. Speech Recognition for Turkish Broadcast News. Signal Processing and Communications Applications (http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4298741).
- [8] Hasan Palaz et al. 2005. Tren – Turkish Speech Recognition Platform. In Proceedings of Eusipco-2005: September 4–8, Antalya, Turkey.
- [9] Nirenburg S., Somers H., Wilks Y. 2003. Readings in Machine Translation. The MIT Press, 413 p.
- [10] Trujillo A. 1999. Translation Engines: Techniques for Machine Translation. Springer-Verlag, 303 p.
- [11] Arnold D. et al. 1996. Machine Translation: an Introductory Guide. London: Blackwell Ltd, 219 p. (<http://clwww.essex.ac.uk/Mtbook>).
- [12] Streiter O., Iomdin L., Hong M. and Hauck U. 1999. Learning, Forgetting and Remembering: Statistical Support for Rule-Based MT. In Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI99), August 23–25, Chester, England.
- [13] Hassan Sawaf, Braddock Gaskill, Michael Veronis. 2008. Hybrid Machine Translation Applied to Media Monitoring. The Eighth Conference of the Association for Machine Translation in the Americas Waikiki, Hawai'i, 21–25 October.
- [14] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. Computational Linguistics, 19:2, pp. 263–311.
- [15] F. J. Och and H. Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, 30: pp. 417–449.
- [16] Cicekli I., Güvenir A. 2003. Learning Translation Templates from Bilingual Translation Examples, in: Recent Advances in Example-Based Machine Translation, Carl, M., and Way, A. (eds), The Kluwer Academic Publishers, Boston, pp: 247–278.
- [17] Cicekli I. 2001. A Specific Least General Generalization of Strings and Its Application to Example-Based Machine Translation, in: Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN2001), North Cyprus, pp: 228–237.
- [18] Altıntaş K., Çiçekli İ. 2001. A Morphological Analyzer for Crimean Tatar. In: Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN2001, pp. 180–189, North Cyprus.
- [19] Cüneyd Tantı, Eşref Adalı and Kemal Oflazer. 2007. A MT System from Turkmen to Turkish Employing Finite State and Statistical Methods, in Proceedings of MT Summit XI.
- [20] Cüneyd Tantı, Eşref Adalı and Kemal Oflazer. 2007. Machine Translation between Turkic Languages, in Proceedings of ACL 2007 – Companion Volume, Prague, Czech Republic, June.
- [21] Iskhakova Kh.F. 1968. Avtomaticheskii sintez form sushestvitelnogo v tatarskom yazike. Sovetskaya tyurkologiya, 2(8): 20–27.
- [22] Pines V.Y. 1974. Nekotore voprosi avtomaticheskogo perevoda i tyurkskie yaziki. Sovetskaya tyurkologiya, 3: 100–107.
- [23] Bektayev K. 1990. Statistika kazakhskogo teksta. Gilim, Almaati.
- [24] Mahmudov M. 2002. Metnlerin formal tehlili sistemi. Elm, Baku.
- [25] Abdullayev A, Seyidov Y, Hasanov A. 1972. Müasir Azərbaycan dili (Modern Azerbaijani language). Maarif, Baku.
- [26] Abbasov A, Fatullayev A. 2007. The use of syntactical and semantic valences of the verb for formal delimitation of verb word phrases. Proceedings of the 3rd L&TC'07: 5–7 October, Poznan, Poland.

- [27] Fatullayev R., Abbasov A., Fatullayev A. 2008. Peculiarities of the development of the dictionary for the machine translation system from Azerbaijani. Proceedings of EAMT 2008, September 22 & 23, 2008, Hamburg, Germany. pp. 35–40.
- [28] Fatullayev R., Abbasov A., Fatullayev A. 2008. Dilmanc is the 1st translation system for Azerbaijani. Proceedings of SLTC 2008: November 20 & 21, 2008, Stockholm, Sweden. pp. 63–64.