The Second International Conference "Problems of Cybernetics and Informatics"
September 10-12, 2008, Baku, Azerbaijan. Section #1 "Information and Communication Technologies" www.pci2008.science.az/1/41.pdf

TURKISH-AZERBAIJANI TRANSLATION MODULE OF DILMANC MT SYSTEM

Abulfat Fatullayev[1], Samir Shagavatov[2]

[1]Institute of Linguistics of ANAS, Baku, Azerbaijan, *fabodilmanc@gmail.com*
[2]Institute of Cybernetics of ANAS, Baku, Azerbaijan, *samir@dilmanc.az*

Abstract. The paper is dedicated to the problem of the development of the machine translation (MT) among Turkic languages. Peculiarities of the automation of the translation process among these languages are investigated; the composition parts of the lexical and grammatical information necessary for providing the translation algorithm and the ways of their creation are defined. The use of the created lexical and grammatical databases in the Turkish Azerbaijani translation module of the Dilmanc MT system is explained.

Introduction

The closeness of the languages belonging to Turkic group (Turkish, Azerbaijani, Kazakh, Uzbek, Turkmen, Kyrgyz, Tatar, …) shows itself from two aspects: closeness on lexical level and closeness in grammatical (morphological and syntactic) level. Despite closeness on the grammatical level, some of these languages are also close on their lexical composition (for example, Turkish-Azerbaijani, Kazakh-Kyrgyz), but some are very different. For example, though Azerbaijani and Kazakh are very close in the grammatical level, but lexical differences are so high that it is difficult to understand the oral and written speech. As example, we can show the following text fragment from the newspaper *"Егемен Қазақстан"*(Independent Kazakhstan, *www.egemen.kz/?act=readarticle&id= 3747)* which is very difficult to understand for the Azerbaijani-speaking persons:

> 2007-02-20: ЛАТЫН ӘЛІПБИІ – КЕЛЕШЕКТІҢ КІЛТІ
> филология ғылымдарының докторы, профессор Әлімхан ЖҮНІСБЕКБЕН әңгіме
> – Әлеке, бұл күндері қоғамымызда әліпби ауыстыру мәселесі кеңінен талқылануда. Қилы-қилы ойлар айтылуда. Осыған байланысты тіл білімінің кәнігі маманы және А.Байтұрсынов атындағы Тіл білімі институтының бас ғылыми қызметкері ретінде сіздің де көзқарасыңызды білгіміз келеді.

So, the languages of Turkic group can be divided into subgroups on the lexical closeness. Because Turkic languages that do not belong to the same subgroup have sufficient number of lexical differences, the understanding the Turkic languages from the different subgroups is as difficult as the understanding the languages that do not belong to Turkic group.

For this reason, researches on the development of the MT systems for Turkic languages are being carried out in two directions:

1. Development of the MT systems among languages of Turkic group;
2. Development of the MT systems from/into languages of Turkic group into/from languages that do not belong to this group.

Because the languages of Turkic group are very close languages in the grammatical level, it is easier to create the MT system of the first type than the second type. Despite considerable number of theoretical scientific works on this direction [1-3], for the present the first practical useful MT system among Turkic languages is only developed in the frame of the Dilmanc MT system (Turkish-Azerbaijani MT system, www.dilmanc.az). It is important to note, that the linguistic technologies developed within any software for any of Turkic languages can be also used for the development of the same software for other Turkic languages.

In this article peculiarities of the development of the first type of the MT systems are considered.

Turkish-Azerbaijani MT system

MT systems between closely-related languages are not rare. The Spanish-Catalan, Czech-

Slovak MT system can be shown as the examples [4-5]. Azerbaijani and Turkish are closely related languages in both levels – lexical and grammatical.

"… advantage of translation between closely related languages is its creating a domain of interchangeable languages. In other words, having a system that is capable of successfully translating between Turkish and Uzbek, any machine translation system translating from English to Turkish will also enable us to translate from English to Uzbek. Implementing a system translating from Turkish to Uzbek is easier than developing a system translating from English to Uzbek. So, with lesser effort, we can have a system that is capable of translating from English to several Turkic languages" [1].

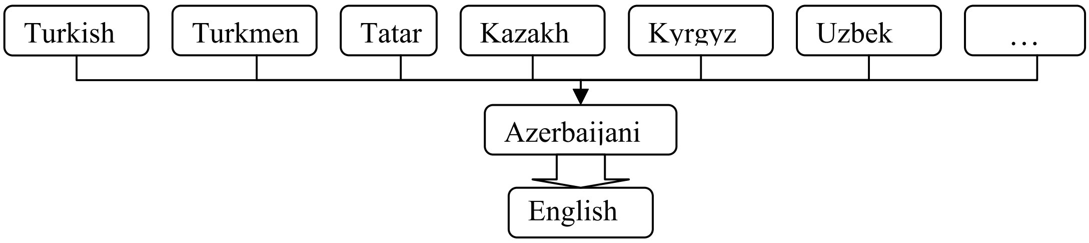
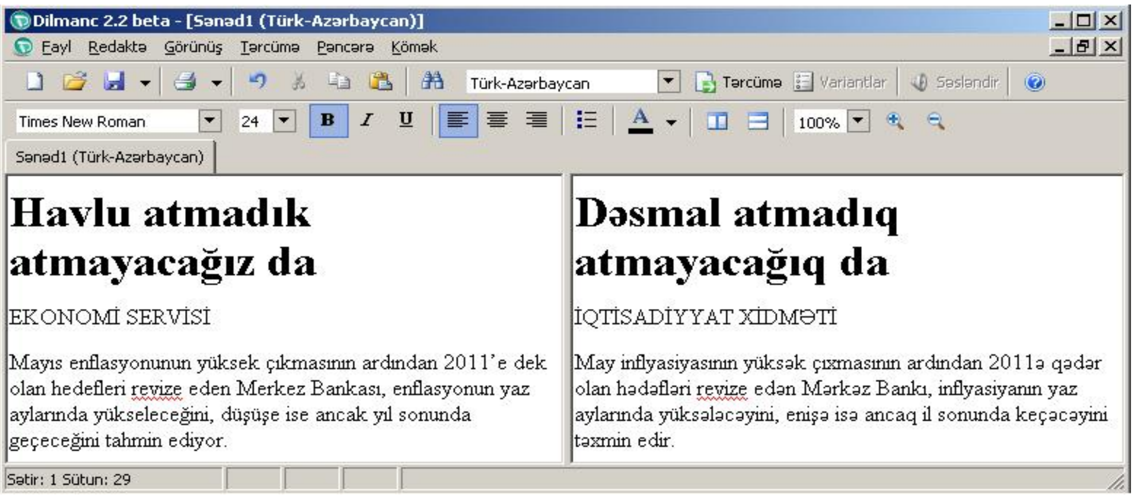


Fig. 1

For this reason, creation of the Turkish-Azerbaijani MT system yields the possibility to develop the MT systems for all languages of Turkic group by using Azerbaijani-English MT technologies developed within the Dilmanc MT system (Fig. 2).



Havlu atmadık atmayacağız da

EKONOMİ SERVİSİ

Mayıs enflasyonunun yüksek çıkmasının ardından 2011'e dek olan hedefleri revize eden Merkez Bankası, enflasyonun yaz aylarında yükseleceğini, düşüşe ise ancak yıl sonunda geçeceğini tahmin ediyor.

Dəsmal atmadıq atmayacağıq da

İQTİSADİYYAT XİDMƏTİ

May inflyasiyasının yüksək çıxmasının ardından 2011ə qədər olan hədəfləri revize edən Mərkəz Bankı, inflyasiyanın yaz aylarında yüksələcəyini, enişə isə ancaq il sonunda keçəcəyini təxmin edir.

Fig. 2. Turkish-Azerbaijani translation module of the Dilmanc MT system

Turkish-Azerbaijani MT system is based on the word-by-word translation approach. This approach can be realized by the creation of the Turkish-Azerbaijani MT dictionary consisting of all word-forms of Turkish. But Turkish as all languages of Turkic group has an agglutinative nature. By adding various suffixes to the stem of the same verb, it is possible to create 17947 word-forms in the Tatar language, 11390 in Turkish and 13592 in Uzbek. In Azerbaijani, the number of word-forms formed from the same stem is more than 8000 [7, p. 84].

The great number of the word-forms leads to the necessity of the finding an alternate way for the creation of the MT dictionary.

This problem is characteristic for both type of MT systems mentioned above. For closely related Turkic languages, it is not so difficult to avoid this problem. Existence of the equivalent suffixes in both – source and target languages leads to the possibility to get the good enough translations in the most cases by replacing the stem and suffixes in the source sentence with their equivalent stem and suffixes in target language.

To this effect, Turkish-Azerbaijani dictionary of stems (Table 1) and the database of the equivalency of the Turkish and Azerbaijani suffixes (Table 2) are created. Except stems and suffixes some other necessary information for the providing the correctness of the translation process is also included in these databases. The small fragments of these databases are shown below.

Table 1. Turkish-Azerbaijani MT dictionary

Turkish	Azerbaijani	Code
abart	şişirt	001
soyut	mücərrəd	004
aday	namizəd	002
yaz	yay	002
yaz	yaz	001
...		

In the 3rd column of this table is shown the special codes instead of the part of speech of the stems (Table 3). It is possible to get acquaintance about the destination of these codes in [6].

The database of the equivalency of Turkish and Azerbaijani suffixes is shown in the Table 2 (Note, we name as suffix both simple and compound suffixes). In this table except the "translations" of Turkish suffixes is also included in some additional information for the correct generation of Azerbaijani word-forms (6th -10th columns).

Table 2. Database of the equivalency of Turkish and Azerbaijani suffixes

Turkish suffix	Azerbaijani equivalent suffix				Type of stem	Connecting consonant	Code of part of speech	Previous letter	Direct connection
	Var. 1	Var. 2	Var. 3	Var. 4					
1	2	3	4	5	6	7	8	9	10
dikleri	dıqları	diklәri	duqları	düklәri	f		004		
e	a	ә			a	y		c	
ca	ca	cә			004		005		
ca	tәrәfindәn				002		005		
sız	sız	siz	suz	süz	a		004		
yacak	acaq	әcәk			f	y		v	
sın	sın	sin	sun	sün	f				*
sın	san	sәn			f				
sın	san	sәn			a				
dıysanız	dınızsa	dinizsә	dunuzsa	dünüzsә	f				
...									

The size of the article does not lead to the possibility to describe in detail of this table, though brief structure of this table is described below.

1st column. Suffix in Turkish;
2nd -5th column. Azerbaijani equivalent variants of the Turkish suffix;
6th column. In this column letter "a" means: stem of the word-form belongs to nominative group (noun, adjective, pronoun, numeral); "f" - verb; Digital code - stem belonging to the part of speech presented by this code.
7th column. Connecting consonant;
8th column. Digital code in this column means that after connection with suffix what part of speech received word-form will belong;
9th column. The type of the letter (v - vowel or c - consonant) in Turkish word-form;
10th column. Suffix is joined to the stem directly.

Using these tables Turkish-Azerbaijani module of the Dilmanc MT system translates Turkish text to Azerbaijani text in the following sequence:

1. Separating Turkish text into sentences and the sentences into the word-forms;

2. Morphological analysis of Turkish word-forms;
3. Lexical and grammatical disambiguation;
4. Definition of Azerbaijani equivalent of Turkish suffixes on the basis of grammatical rules of Azerbaijani (vowel harmony rule, sequences of simple suffixes in suffix chain etc.);
5. Synthesis of Azerbaijani translations of Turkish word-forms;
6. Generation of Azerbaijani sentences and text.

Table 3. Codifying system

Code	Part of speech
001	Fel (Verb)
002	İsim (Noun)
003	Əvəzlik (Pronoun)
004	Sifət (Adjective)
005	Zərf (Adverb)
006	Say (Numeral)
...	

Example. Let's consider the morphological analysis process of the Turkish word-form *yazdıysanız*. First the stem and suffix chain of the word-form is separated – *yaz* and *-diysanız*. This stem has two translations in Azerbaijani – *yaz* and *yay* and the suffix *–diysanız* has 4 equivalent variants *–dınızsa, –dinizsə, –dunuzsa, –dünüzsə* (Table 2). On the vowel harmony rule we take *–dınızsa* variant in Azerbaijani. While translating into Azerbaijani we can construct two word-forms which have the different meanings – *yay-dınızsa* and *yaz–dınızsa.* Because the suffix *–diysanız* is a suffix which can be jointed to only verb stem (Table 2, column 6) we take the second record for the stem *yaz* and consequently receive the Azerbaijani word-form *yazdınızsa.*

Conclusion and future works

Turkish-Azerbaijani MT system is the first practical useful system between two members of the Turkic group. For the present, there are about 20000 stems and 1000 suffixes in the database of this MT system. Despite some lexical and grammatical ambiguity solving problems the translation module gives the good enough translations of Turkish texts into Azerbaijani. Further the volume of Turkish-Azerbaijani dictionary will be increased and the translation algorithms will be improved. It is possible to improve the technologies developed within this project and also to apply to the creation of the MT systems among other languages of Turkic group.

The researches of the authors on the development of the MT systems among languages of Turkic group are going on.

References

[1] Altintas K., Cicekli I. A Machine Translation System between a Pair of Closely Related Languages, in: Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002), Orlando, Florida, 2002, CRC Press, pp. 192-196.

[2] Cüneyd Tantuğ, Eşref Adalı and Kemal Oflazer, A MT System from Turkmen to Turkish Employing Finite State and Statistical Methods, in Proceedings of MT Summit XI, 2007.

[3] Tantuğ C., Adalı E., Oflazer K. Machine Translation between Turkic Languages, in Proceedings of ACL 2007 – Companion Volume, Prague, Czech Republic, June 2007.

[4] Canals-Marote R. et al. Pastor-interNOSTRUM: a Spanish-Catalan Machine Translation System. Machine Translation Review, No.11, December 2000, pp. 21-25.

[5] Kubon V., Hajic J., Hric J. 2000, Machine Translation of Very Close Languages, in ANLP-NAACL2000, Washington (www.bilkend.cs.edu.tr)

[6] Fatullayev A. Development of the digital method for the Azerbaijani-English MT system and its application. Namizədlik dissertasiyası, AMEA Kibernetika institutunun kitabxanası

[7] Mahmudov M. Mətnlərin formal təhlili sistemi. Bakı, Elm, 2002, 259 p.