



COMPUTATIONAL LINGUISTICS AND RESEARCHES ABOUT UYGHUR LANGUAGE

Hesaplamalı Dil Bilimleri ve Uyghur Dili Araştırmaları

Murat ORHUN*

Abstract

In this paper computational linguistics explained briefly and recent computational linguistic researches about Uyghur Language are summarized. With developing computer technology, computer based researches about different languages have been made a big progress. For example, content management of a text, information retrieval, dialog systems, document clustering, text mining, spell checking, speech to text conversion, text to speech conversion and automatic translation between different languages (machine translation) systems have been developed and used in real life. Though there are so many computational researches have been done about some languages such as Finish, Japanese, Hungarian and Turkish that member of Ural-Altai languages family, but the researches about some other languages, for example such as Uyghur language, are less known. To improve computational researches and analyze relations between different languages, latest computational researches about Uyghur language, especially fundamental works that related to machine translation are summarized in this article mainly. At the same time, relations between linguistics and computational linguistics are analyzed.

Keywords: Uyghur language, Uyghur morphology, computational linguistics, machine translation, Uyghur corpus.

Özet

Bu makalede hesaplamalı dil bilimleri kısaca anlatılmıştır ve Uyghurca ile ilgili yapılan güncel hesaplamalı dil bilim araştırmaları özetlenmiştir. Teknolojinin ilerlemesi ile farklı dillere yönelik bilgisayar destekli çalışmalarda büyük başarılar elde edilmiştir. Örneğin, metinlerde içerik yönetme, bilgi edinme, konuşma sistemleri, dosya kümeleme, metin madenciliği, yazı kontrolü, yazıyı sese çevirme, sesi yazıya çevirme ve farklı diller arasında otomatik (bilgisayarlı çeviri) gibi uygulamalar geliştirilmiştir ve gerçek hayata kullanılmaktadır. Gerçi Fince, Japonca, Macarca ve Türkçe gibi Ural-Altay dilleri grubuna ait bazı diller ile ilgili birçok çalışmalar yapılsa bile, ancak yine bazı diller, örneğin Uyghurca, ile ilgili yapılan çalışmalar çok az bilinmektedir. Hesaplamalı dil bilimi ile ilgili araştırmaları geliştirmek ve farklı diller arasındaki ilişkileri analiz edebilmek için, bu makalede, Uyghurca ile ilgili yapılan bilgisayar destekli araştırmalar, özellikle ile bilgisayarlı çeviri ile ilgili yapılan en son temel niteliğindeki çalışmalar toparlanmıştır. Aynı anda dil bilimcileri ile hesaplamalı dil bilimleri arasındaki bağıntı analiz edilmiştir.

Anahtar Kelimeler: Uyghur dili, Uyghurca biçim bilgisi, hesaplamalı dil bilimleri, bilgisayarlı çeviri, Uyghur derlem.

*Assist. Prof. Dr., Computer Engineering Department, Faculty of Engineering and Natural Science, Istanbul Bilgi University, Istanbul, TURKEY. E-mail: murat.orphun@bilgi.edu.tr.

Introduction

In science, the field that studies languages with computer science is called computational linguistics. Computational Linguistics consists of two components such as computer science and linguistics. The main contribution of linguistics in computational linguistics are, analyze all property of a natural language without making any constraints. It means don't change the natural property of a natural language. In this aspect, natural language analyzing is completely different form analyzing formal languages such as computer programming languages. The responsibility of computer science is, design and implement algorithms that solve problems related to a language. There are some different subfields of computational linguistics, for example, speech recognition which deals how a spoken language could be understood by a computer, speech synthesis that studies how to generate a speech with texts, information retrieval that gets main information from a big size of document etc. One of the most popular fields of the computer linguistics is machine (mechanical) translation between different languages. The first most simple machine translation was completed in 1950's. This system used to translate some Russian political sentences into English. However, at that time, the computers were very expensive and the software technology were not developed as today, the expected system was not successful and the famous ALPAC report was published (Hutchins 1995: 439-440). This report made a negative effect on this field, but these researches continued different parts of the world. For example, the translation system RUSLAN that translates from Czech to Russian was developed (Hajic 1997). After this translation system, different translation systems have been implemented such as Czech to Slovak (Hajic 2000), Japanese to English (Nagao 1994), French to English and German to English (Hanneman 2008). The computational researches about Turkic languages so far have been done on the Turkish language. There are many computational research groups in different universities in Turkey and many machine translation systems such as Turkish to Azerbaijani (Hamzaoğlu 1993), Turkish to Crimean Tatar (Altıntaş 2000) and Turkmen to Turkish (Tantuğ 2007) have been implemented. The main properties of these translation systems about the Turkic languages are, they all based the same technology, two level morphological analyzer that developed for Turkish (Koskeniemi 1983). For implementation of the morphological analyzer of the Crimean Tatar (Altıntaş 2001) and Turkmen language (Tantuğ 2006), the same tags used that used to implement Turkish morphological analyzer (Oflazer 1995).

The computational researches about Uyghur language are still initiating stage and this paper summarizes main computational researches of the Uyghur language. The rest of the paper is organized as follows. The next section will introduce Uyghur languages and Uyghur alphabets. Section three introduces morphological analyzer for Uyghur languages. Section four introduces morphological disambiguator. Section five introduces the development of Uyghur corpus. Section six introduces the proposed machine translation systems from Uyghur to Turkish. Section seven gives a conclusion about computational researches about Uyghur language.

Uyghur Language and Alphabets

The Uyghur language is a Turkic language that belongs to Ural- Altaic language family. The Uyghur language is spoken mainly by Uyghur people that live in Xinjiang, the North West part of China. Except this there are some Uyghur people live in Central Asian Turkic republics, Turkey, Europe and USA. In history, Uyghur people invented different alphabets (Orhun-Yenisey, Old Uyghur, Chagatai, and Latin) and used them at different period of times (Kaşgarlı 1992: 41-57). At present Uyghur people are using the Arabic alphabet officially in Xinjiang. They have been using this alphabet since 1983 (Kaşgarlı 1992:44). In Uyghur literature there are many important publications have been published with these alphabets such as Latin and Arabic scripts (Ayup 2014). In some case, to understand

previous published literature in different alphabets, it is necessary to understand old scripts as well (Abdulla 2016). As a result of the alphabet reform happened in recent history, there are lot of people prefer to use the Latin alphabet unofficially that used before the Arabic Alphabet. Therefore most of the web pages are published both in Arabic and Latin scripts. Unfortunately there is not an official organization to standardize alphabets for computers, different software companies developed their own encoding system for characters. In order to solve these problems and standardize both of Arabic and Latin characters for computer, an extensive computational researches have been done (Duval 2006).

With the availability of the computers, especially Internet, the Latin alphabet is getting popular even before. Most of web sites are publish with two different alphabets such as, Arabic, Latin and Cyrillic¹. Due to none of the web developers get the formal training about Latin alphabet, there were some confusion happened about double characters. In order to solve this problem, the Uyghur Computer Science Association² (UKIJ³) is formed by some volunteer researchers. This association coded all of the Arabic characters with Unicode first and then implemented a software that translates Arabic characters to modified Latin script. This software distributed freely through the Internet and made a big contribution for Uyghur language. In this way, all of the documents began to be written with these characters and this alphabet become standard alphabet for Uyghur language. The comparison between Arabic and Latin scripts is given in Figure 1. In this new Latin script; some double characters are used to represent some Arabic characters.

ئا	ئە	ب	پ	ت	ج	چ	خ	د	ر	ز	ژ	س	ش	غ	ف
a	e	b	p	t	j	ch	x	d	r	z	j/zh	s	sh	gh	f
ق	ك	گ	ڭ	ل	م	ن	ھ	ئو	ئۆ	ئۇ	ۋ	ئې	ئى	ي	ى
q	k	g	ng	l	m	n	h	o	u	ö	ü	w	é	i	y

Arabic-Script Uyghur alphabet and its unified correspondence in Latin-Script (Duval 2006).

Uyghur language is an agglutinative language and new words can be created with adding suffix to a root word. After a word has been added to a root word, the property of the root word may be changed according to added suffix. After a root word has been added an suffix, A new word will be created and this created word can be added by another suffixes. For example⁴: one of Uyghur root word and some of possible generated words are given in Table I.

Table 1: Uyghur words that generated from the root word “ish”

¹ Cyrillic is the official alphabet in Kazakhstan and Kyrgyzstan republics. Cyrillic alphabet is not explained in this article.

² It is a nongovernmental organization supported by linguists and software engineers. This organization makes software with GPL license and supports to develop software technology for Uyghur and other Turkic people in the region.

³ UKIJ: abbreviation of Uyghur Computer Science Association in Uyghur language of, “Uyghur Kompyuter Ilimi Jemiyiti (UKIJ)”. <http://www.ukij.org/fonts/>

⁴ All of the examples in this paper are written in Latin Uyghur scripts.

	Uyghur	English
1	ish	work, task, job
2	ish+chi	worker
3	ish+xana	office
4	ish+le	work
	ish+le+me	Do not work
5	ish+siz	jobless
6	ish+chi+lar+ning+ki+mu	Is the things that belong to workers ?
7	ish+chi+lir+imiz+ning+ki+dek	it seems that look like to our workers'

In order to analyze a word with a computer, all of the words must be recognized by a computer first. It means all words in a dictionary or in a language must be saved in a computer memory. Theoretically, there are millions of words can be generated with adding affixes to a root word (Oflazer 1995: 138). As a result, the most efficient way to implement a dictionary is, design it that consists up with root words only and represent all affixes and suffixes with related tags. After that design rules to connect root words and suffixes according language grammar.

In previous example the word “ish” (work) can be analyzed as: the root of the word is, “ish” it is a noun, in singular form, doesn’t have personal suffix and doesn’t have any case suffixes.

In the same way, the word on the second row, “ish+chi” (worker) can be analyzed as: the root of the word is, “ish”, it is noun, in singular form, doesn’t have personal affix and doesn’t have case suffixes either. But there is one more affix here, “chi”. In Uyghur language the “chi” suffix generates a noun when followed a noun (Tömür 1997: 85, Osmanof 1997: 906, Öztürk 1993: 25). Also, the “chi” suffix creates an adjective if it follows a noun or an adjective (Tömür 1997: 119). In this way, instead of saving two words such as “ish” and “ish+chi”, just save the root word “ish” in a dictionary only.

In order to analyze Uyghur words with a computer, a morphological analyzer have been implemented with the finite states automate technology (Orhun 2009a; Orhun 2009b). In this morphological analyzer there are 600 root words and 100 affixes are used. Also there are 30 more two-level rules are used to simply the analyzer (Orhun 2010). Because the root words are limited in this morphological analyzer, therefore it is not possible to analyze all words in the contemporary Uyghur language. Also this morphological analyzer could analyze basic Uyghur words, unfortunately cannot analyze adapted words that taken from other languages such as, Chinese, Russian, Persian ant Arabic etc.

Even though this morphological analyzer is still under developing stage, words that recorded its dictionary could be analyzed. For example, with this analyzer, the word “ish” can be analyzed as:

ish: ish+Noun+A3sg+Pnon+Nom

In the solution, the “+” sing represents the morpheme borders of a word. The solution

could be read as: the root of the word is a noun (Noun), it is in singular form (A3sg), doesn't take personal suffixes (Pnon), and doesn't take case suffix (Nom) either. If the word "ishchilarni" is analyzed, the following solution could be generated:

iş+Noun+A3sg+Pnon+Nom^DB+Noun+Agt+A3pl+Pnon+Acc

This solution could be read as: the root of the word is "ish", it is a noun (Noun), in singular form (A3sg), doesn't have personal suffix (Pnon), doesn't have case suffix (Nom), a new type word is generated (+DB), the type of the new word is, noun (Noun), the suffix is (Agt)⁵, this noun is in plural form (A3pl), doesn't have personal suffix but it has the accusative suffix (Acc).

Morphological analyzer is one of the most important works for any kind language. Especially for agglutinative languages, the first thing is implement a morphological analyzer. Because, with a morphological analyzer, all of the properties of a word could be analyzed in detail. In general, a morphological analyzer is implemented according to a specific languages grammar. So a morphological analyzer can be used as word recognized of a language at the same time. If an invalid word tried to analyze, no solution will be generated.

There are two possibilities for this case if no solutions is generated, may be the related root word is not found in the dictionary, or the word is constructed not grammatically. For example: if the word, "kitap+ni+lar" submit to Uyghur morphological analyzer, there are no results will output. Because, in Uyghur language, no affixes are followed after the accusative suffix "ni". So this word "kitap+ni+lar" is an invalid word. The correct form of this word is: "kitap+lar+ni" Most of the spell checking and correcting algorithms work on this principle.

Morphological Disambiguator

In general, we prefer to get as many information as possible about a word with a morphological analyzer. But if the next approach is about machine translation, then we need to use only one solution. The task selection the correct morphological solution of a word is called morphological disambiguation.

For example the word "at" will be generate the following solution if analyzed with the current Uyghur morphological analyzer:

at+Noun+A3sg+Pnon+Nom

at+Verb+Pos+Imp+A2sg

The first solutions means, the root word is "at" and it is a "noun", but the second solution means, the root of the word is "at" and it is a "verb". Suppose we are going to implement a machine translation system, so which solution we should use? Because if use "at" as a noun, then we should translate such as, "ad"(name), or "at"(horse) and if we translate it as a verb, then we need to translate it such as "throw". So the morphological disambiguation is really very important and crucial task of the natural language processing.

In general there are two kinds of solutions are used, corpus based solution and rule based solution. Both of the solutions are related to a context of a word that occurs in it. For Uyghur language, both of the solutions have been implemented recently. In the corpus based solution, there is a raw Uyghur corpus have been used that consists up with 594,192 words. Then this corpus is divided into some sub corpus in order to calculate some specific word categories (Aisha 2009a, Aisha 2009b).

⁵ In this morphological analyzer, the "Agt" tag is used to mark the "chi" suffix.

For rule based system, some specific rules are defined according to the Uyghur grammar (Orhun 2010). With this rule, some general ambiguities about a word could be solved. For example: the sentence, “bu yaz issiq” (this summer is hot) will be generates the following solutions when analyzed with the morphological analyzer:

bu: bu+Pron+Demons+A3sg+Pnon+Nom

yaz: yaz+Noun+A3sg+Pnon+Nom*

yaz: yaz+Verb+Pos+Imp+A2sg

issiq: issiq+Adj

As a result there are two solutions are generated about the “yaz” word. In Uyghur languages the “yaz” will be translated as “summer” when it solved as noun, and will be translated as “write” if solved as a “verb”. In this sentence the “yaz” should be solved as a noun according to its context. In this sentence, the rule to select the correct solution is related to a verb. In Uyghur language, an imperative verb always occurs end of a sentence. But in this sentence it is occurred middle of a sentence (before the adjective “issiq”), so the solution that related to a “verb” is not valid. In this way the correct solution related to a “noun” is selected correctly.

Another example, “sen kitap yaz” (write a book), are will be generated following solutions when analyzed:

sen+Pron+Pers+A2sg+Pnon+Nom

kitap+Noun+A3sg+Pnon+Nom

yaz: yaz+Noun+A3sg+Pnon+Nom

yaz: yaz+Verb+Pos+Imp+A2sg

In these solutions, there is only one solution for, “sen” and “kitap” except, “yaz”. If there is only one solution for a word, then that solution could be accepted as a valid solution directly. Unfortunately, there are so many ambiguous words in Uyghur languages and the percentage is 1.44 solutions for per word (Orhun 2010). In this sentence, the rule to select the solution is between a “verb” and a “noun” that occur end of a sentence. In Uyghur language, in most of the cases, verbs occur end of a sentence. Because of this reason, we can select the solution that solved as a verb. Also, one more reason that we select the solution which is about a verb is, the verb is in “imperative” mood, tagged with “Imp”. So all of the imperative verbs definitely occur end of sentence.

The problem about rule-based solution is, it is necessary to detect all of the cases. As a natural language there are many rules exist and it is difficult to define all of them. For corpus base solution, the correct solution could be selected if the corpus has enough data.

Corpus

In linguistics, a corpus or text corpora is a large and structured set of texts. They are used to statically analyzing words automatically. With increasing electronic documents, especially availability of the Internet, the corpus based linguistics systems getting more popular than rule based systems. Because, if a good implemented corpus available, then most problems about a language such as, morphological ambiguity and word sense ambiguity could be solved without defining any rules. For example the WordNet (Gorge 1995) corpus was begun to construct in 1985 and supported with big amount of funds for English. Later for English then extended to other languages.

In general, a corpus consists up with a large set of data, for example, the WordNet⁶ corpus has 207,000 word- sense pairs in total. In usual, the annotated corpus are preferred but their drawbacks are, about their sizes. Because in an annotated corpora all of the sentences must be solves according to a language grammar. Also it must be correct. So it needs very professional linguistic knowledge. Most of the time annotated corpus are referred as TreeBank corpora (Abeille 2003).

For Uyghur language, the beginning of constructing the Uyghur corpus was began in 2001. After that many improvements have been made. One of the first any systematic development of the Uyghur corpus is, the tagged Uyghur corpus that was implemented by Yusup and Lee (Aibaidula 2003). In this corpus there are 8 million words used that selected from daily newspaper, school books (from elementary to university), economics, history, religion etc. In this corpus there are 60,000 root words and 1300 suffixes (terminal) were used. There are 12 main tags and 50 sub tags were used to annotate corpus words. For example, the “noun” category is represented with “N”. If the noun is related to a “time” then it is represented with “Nw”. Also the are some rules defined according to Uyghur grammar. All of the words are tagged according to statistical results and defined rules. Based on this corpus, there are many knowledge based application have been implemented such as word processing, phrase processing etc. (Aibaidula 2009).

Machine translation between Turkic languages

All of the Turkic languages have the same syntax. This is one of the main advantages for machine translation systems between different languages. All of the Turkic languages compared syntactically and morphologically in detail by Tantuğ (Tantuğ 2007). After his research he proposed a hybrid system that translates between all of the Turkic languages (Tantuğ 2007). Also he implemented a machine translation system between Turkmen and Turkish (Tantuğ 2007b).

From his view, a machine translation system has been implemented between Uyghur and Turkish (Orhun 2011). This system translates simple sentences from Uyghur to Turkish. For example, the Uyghur sentence “men kitap yazdim” (I wrote a book) can be translated into Turkish with the following process. First the source sentence will be analyzed with the Uyghur morphological analyzer such as:

men: men+Pron+Pers+A1sg+Pnon+Nom

kitap: gosh+Noun+A3sg+Pnon+Nom

yazdim: yaz+Verb+Pos+Past+A1sg

After each of the words is analyzed, then the root words are translated according to a bilingual dictionary:

man -> ben

kitap -> kitap

yaz-> yaz

After the root words have been translated in to Turkish, the root words are combined with the morphemes (suppose there is not morphological ambiguity).

ben: men+Pron+Pers+A1sg+Pnon+Nom

⁶ WordNet: This corpus is improving continuously, see : <http://wndomains.fbk.eu/>

kitap: gosh+Noun+A3sg+Pnon+Nom

yaz: yaz+Verb+Pos+Past+A1sg

After this morphological root words and morphemes have been submitted to the morphological generator of the Turkish language (Oflaz 1995), and then the surface sentence will be generated as:

“ben kitap aldım”

The main difference between Uyghur to Turkish and Turkmen to Turkish is, there are disambiguation rules used in Uyghur to Turkish translation system. In this system a word-sense disambiguation system is implemented in order to improve translation quality. Though both of the Uyghur and Turkish language have the same sentence structure, still there are some differences related to morphemes (Orhun 2009c).

Conclusion

In this paper most recent computer based researches about Uyghur languages have been summarized. Actually some rule-based and corpus based machine translation systems have been implemented for Uyghur language. For example, translation systems from Uyghur to Japanese (Mahsut 2004, Nimaiti 2012) and Uyghur to Chinese (Yang 2009, Dong 2010). As we know in natural language processing, the constructing corpus is very important field and most of the recent researches are based corpus. Even the Uyghur corpus have been constructed, it is still not available through the world for educational purpose. In case a researcher wants to do a research for Uyghur languages, especially intended methods related statistical calculation, the first thing is to implement a test corpus (Aisha 2009a, Aisha 2009b, Belkiz 2007a, Belkiz 2007b; Eziz 2007). Because of different tags are used in these corpus, it is impossible to test different researchers methods. In Turkey, all of the natural language processing researches are based in the same tags that used in Turkish Treebank corpora (Atalay 2003). Therefore, if all researchers use the same tags in their researches for different Turkic languages, it is possible to develop a system that works together.

References

- ABAI DULLA Yusup et al.(2009), “Progress on Construction Technology of Uyghur Knowledge Base”, International Symposium on Intelligent Ubiquitous Computing and Education.
- ABDULLA Ayshemgul (2016), “Chaghatay Language is The Bridge Between Old Turkic and Modern Uyghur Language”, International journal of Uyghur Studies, No: 7: Pages: 1-7.
- ABEILLE Anne (2003), “Treebanks: Building and Using Parsed Corpora”, Dordrecht: Kluwer Academic Publishers.
- AIBAI DULLA Yusup; Lua Kim-Teng (2003), “The development of Tagged Uyghur Corpus”, Proceedings of PACLIC17, Sentosa, Singapore, Pages: 228-234.
- AISHA Batuer; Maosong Sun (2009a), “A Uyghur Morphem Analysis Method Based on CONDITIONAL Random Fields”, International Journal on Asian Language Processing, 19(2), Pages: 69-84.
- AISHA Batuer; Maosong Sun (2009b), “A Statical Method for Uyghur Tokenization”, In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, Pages: 1-5.

- ALTINTAŞ Kemal; Çiçekli, İlyas (2001), "A Morphological Analyzer for Crimean Tatar", in Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN North Cyprus, Pages: 180-189.
- ALTINTAŞ Kemal (2000), Turkish to Crimean Tatar Machine Translation System, Computer Engineering Department. Msc. Thesis. Bilkent University, Ankara.
- ATALAY B. Nart et al.(2003), "The Annotation Process in Turkish Treebank", Proceedings of the EACL Workshop on Linguistically Interpreted Corpora.
- AYUP Tursun (2014), "The Development of Turkic Studies in China in the Last 20 Years", International journal of Uyghur Studies, No: 4. Pages: 1-10.
- Belikiz (2007a), "The 3253 different word forms of Uygur Verb /qil/ ", Corpus Linguistics and Corpus Based Research Dept. of Linguistics, College of Anthropology Xin Jiang Normal University, <http://www.xjcorpus.net>
- Belikiz et al. (2007b), "The 2107 different words forms of Uyghur verb /bol/ ", Corpus Linguistics and Corpus Based Research Dept. of Linguistics, College of Anthropology Xinjiang Normal University, <http://www.xjcorpus.net>.
- DONG Xinghua et al. (2010), "Chinese-Uyghur Statistical Machine Translation: The Initial Explorations", The 4th International Universal Communication Symposium (IUCS). Beijing, China.
- DUVAL J.Rahman; JANBAZ, W. Abdulkerim (2006). "An Introduction to Latin Script Uyghur", Middle East & Central Asia Politics, Economics, and Society Conference. Sept 7-9, University of Utah, Salt Lake City, USA.
- EZIZ Gülnar (2007), "Resistance to Borrowing of Uyghur Verbs", Annual Conference, University of Washington, October 18-21.
- George A. Miller (1995), "WordNet: a lexical database for English", Communication of the ACM. Volume 38, Issue 11. Pages: 39-41
- HAJIC Jan (1987), "RUSLAN - An MT System Between Closely Related Languages", in Third Conference of the European Chapter of the Association for Computational Linguistics (EACL'87) Copenhagen, Denmark. Pages: 113-117
- HAJIC Jan et al. (2000), "Machine translation of very close languages", in Proceedings of the Sixth conference on Applied natural language processing, Morgan Kaufmann Publishers Inc. Pages: 7-12.
- HAMZAOĞLU İlker (1993), Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages, in Institute for Graduate Studies in Science and Engineering. MSc Thesis, Bogazici University, Istanbul.
- HANNEMAN Greg et al. (2008), "A Statistical Transfer Systems for French-English and German-English Machine Translation", Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, Pages: 163- 166.
- HUTCHINS W. John (1995), "Machine Translation: A Brief History, Concise history of the language sciences: from the Sumerians to the cognitivists", Edited by E. F. K. Koerner and R. E. Asher, Oxford: Pergamon Press, Pages: 431- 445.
- KAŞGARLI S. Mahmut (1992), Modern Uygur Türkçesi Grameri, İstanbul Orkun Yayınevi.
- KOSKENNIEMI Kimmo (1983), "Two-level morphology: A general computational model for word form recognition and production", Publication No: 11, Department of General Linguistics, University of Helsinki.

- MAHSUT Muhtar et al. (2004), "An Experiment on Japanese-Uighur Machine Translation and Its Evaluation", Conference of the Association for Machine Translation in the Americas AMTA 2004: Machine Translation: From Real Users to Research, Pages: 208-216.
- NAGAO Makoto (1984), "A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle", Proc. Of the International NATO symposium on Artificial and Human Intelligence, Lyon, France, Pages: 173-180
- NIMAITI Maimitili; Yamamoto, Izumi (2012). "A Rule Based Approach for Japanese- Uighur Machine Translation System", Cognitive Informatics & Cognitive Computing (ICCI*CC), IEEE 11th International Conference, Kyoto, Japan.
- OFLAZER Kemal (1995), "Two-level Description of Turkish Morphology", Literary and Linguistic Computing, Vol. 9, Pages: 137-148.
- ORHUN Murat et al. (2009a), "Rule Based Analysis of the Uyghur Nouns", International Journal of Asian Language Processing, 19(1), Pages: 33-43.
- ORHUN Murat et al. (2009c), "Computational comparison of the Uyghur and Turkish Grammar", The 2nd IEEE International Conference on Computer Science and Information Technology, Beijing, China.
- ORHUN Murat et al. (2010), "Morphological Disambiguation Rules for Uyghur Language", IEEE International Conference on Software Engineering and Service Science (ICSESS 2010), July 16-18, Bei Jing, China.
- ORHUN Murat et al. (2011), "Uygurcadan Türkçeye bilgisayarlı çeviri", İstanbul Teknik Üniversitesi, Mühendislik Dergisi, Cilt 10, Sayı 3, Pages 3-14.
- ORHUN Murat et al.(2009b), "Rule Based Tagging of the Uyghur Verbs", Fourth International Conference on Intelligent Computing and Information Systems, Faculty of Computer & Information Science, Ain Shams University Cairo, Egypt, Pages: 811-816.
- OSMANOF Mirsultan (1997), Hazırqi Zaman Uyghur Edebiy Tilining İmla ve Teleppuz Lughiti. Shin Jiang Xeliq Neshiryati.
- ÖZTÜRK Ridvan (1993), Yeni Uygur Türkçesi Grameri, Türk Dil Kurumu yayınları: 593.
- TANTUĞ A. Cüneyd (2007a), Akraba ve Bitişken Diller Arasında Bilgisayarlı Çeviri İçin Karma Bir Model. Bilgisayar Mühendisliği Bölümü. Doktora Tezi, İstanbul Teknik Üniversitesi, İstanbul.
- TANTUĞ A. Cüneyd et al. (2006), "Computer Analysis of The Turkmen Language Morphology", FinTAL Lecture Notes in Computer Science, Vol.4139, Springer, Pages: 186-193.
- TANTUĞ A. Cüneyd et al. (2007b), "Machine Translation between Turkic Languages", Proceedings of the ACL 2007 Demo and Poster Sessions, Pages: 189-192.
- TÖMÜR Hamit (1997), Modern Uygur Grammar (Morphology), Yıldız Teknik Üniversitesi, Fen-Edebiyat TDE Bölümü. İstanbul.
- YANG Pan et al. (2009), "Chinese-Uyghur Machine Translation System For Phrase-Based Statistical Translation", Journal of Computer Applications 29(7).