itüdergisi/***d***
mühendislik
Cilt: 10, Sayı: 3, 3-14
Haziran 2011

Uygurcadan Türkçeye bilgisayarlı çeviri

Murat ORHUN*, Eşref ADALI, A.Cüneyd TANTUĞ

İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Programı, 34469, Ayazağa, İstanbul

Özet

Bilgisayarlı Çeviri (BÇ) yapay zeka çalışmalarının bir alt dalı olan Doğal Dil İşlemenin (DDİ) alt konusudur. Diller arası çeviride bilgisayarların kullanılması fikri 1950'lerin ilk yıllarında ortaya çıkmıştır. O tarihten günümüze kadar pek çok dil üzerinde çalışılmış ve çeşitli yöntemler geliştirilmiştir. Ancak teknolojideki ve yöntemlerdeki gelişmelere karşın, genel amaçlı, yüksek başarıma sahip çeviri sistemleri henüz geliştirilememiştir. Bunun temel nedeni, diller arasındaki büyük yapısal ve anlatım farklılıklarıdır. Yapısal yönden benzer olan diller arasına bilgisayarlı çevirinin daha kolay olduğu bilinmektedir. Son yıllarda Çekçe-Slovakça, Çekçe-Lehçe, İspanyolca-Katalanca, Türkmence-Türkçe gibi yakın diller arasında yüksek başarımlı çeviri yapabilen sistemler geliştirilmiştir. Akraba veya yakın diler arasında çeviri amaçlı geliştirilen sistemler, farklılıkların büyük olduğu, Türkçe-İngilizce gibi diller arasında bilgisayarlı çeviri için gerek duyulan karmaşık yöntemlere göre, daha basit ve kolay gerçeklenebilir yöntemler kullanmaktadırlar. Bu çalışma kapsamında, aynı dil ailesi içinde sınıflandırılan ve birçok yönden benzerlikler gösteren Uygurcadan Türkçeye bilgisayarlı çeviri sistemi geliştirilmiştir. Aslında bu diller ne kadar benzer özellikler gösterse de, çözülmesi gereken farklılıklar azımsanmayacak kadar çoktur. Genel olarak Uygur Türkçesi ile Türkiye Türkçesinin söz dizimi aynıdır. Bundan dolayı çeviri sistemi geliştirirken, sözcüklerin dizilimi değişmemektedir. Ancak sözcüklere eklenen ekler çok farklılaşabilmektedir. Uygurca ve Türkçe bitişken diller olduğundan, ekler çok önemlidir. Ekler sözcüklerin hatta tümcenin anlamını değiştirmektedir. Bu çalışmada, akraba ve bitişen diller arasında bilgisayarlı çeviri için geliştirilen karma model üzerine, belirsizlik giderme yönteminin eklenmesi ile Uygurcadan Türkçeye bilgisayarlı çeviri sistemi geliştirilmiştir.

Anahtar Kelimeler: *Bilgisayarlı çeviri, Türk Dilleri, Uygurca, Türkçe.*

*Yazışmaların yapılacağı yazar: Murat ORHUN. muratmehmet@cs.bilgi.edu.tr; Tel: (212) 311 53 17.
Bu makale, birinci yazar tarafından İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Programında tamamlanmış olan "Uygurcadan Türkçeye bilgisayarlı çeviri" adlı doktora tezinden hazırlanmıştır. Makale metni 10.05.2010 tarihinde dergiye ulaşmış, 17.06.2010 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 31.08.2011 tarihine kadar dergiye gönderilmelidir.
Bu makaleye "Orhun, M., Adalı, E., Tantuğ, A C., (2011) 'Uygurcadan Türkçeye bilgisayarlı çeviri', İTÜ Dergisi/D Mühendislik, 10: 3, 3-14" şeklinde atıf yapabilirsiniz.

Machine translation from Uyghur to Turkish

Extended abstract

Machine translation is a sub-field of Natural Language Processing which belongs to Artificial Intelligence. Generally, it is based on computer technology that uses software to translate one natural language to another. In the 1950s, the Georgetown experiment involved fully-automatic translation of over sixty Russian sentences into English (Hutchins, 2004). The experiment was a great success and ushered in an era of substantial funding for machine-translation research. One of the main projects initiated by the US at that time was a machine translation system which converted Russian to English.

This project continued from 1950 to 1960. In 1964, government sponsors of machine translation in the United States formed the Automatic Language Processing Advisory Committee (ALPAC) to examine the project's potential. In the famous 1966 report, ALPAC concluded that machine translation was slower, less accurate and twice as expensive as human translation, and that "there is no immediate or predictable prospect of useful machine translation" (Hutchins, 1995). The effects of this report brought about the virtual end to machine translation research in the US for over a decade after its publication.

As computer technology developed, high capacity and high speed computers were produced. Thus, the main restrictions of studying natural language were removed and machine translation gained the attention of the computer science community once again. Despite technologic advances and the advent of new methods, a general purpose for full automatic machine translation systems still does not exist. To date, few machine translation systems have been developed, furthermore, they may only be applied to restricted texts and some post-editing works (usually necessary after initial translations). The main reasons for these are the morphological, syntactical and lexical differences between different languages. In conclusion, translated texts remain inferior to higher quality translations.

Recently, some machine translation systems designed for related languages, such as: Czech to Slovak, Spanish to Catalan, and Turkmen language to Turkish have been implemented; studies on them have proven successful translations can be produced efficiently.

In this study, our aim was to implement a machine translation system between Uyghur language and Turkish. Uyghur language is an agglutinative language such as other Turkic languages (i.e. Turkmen, Kazakh, Kyrgyz, Uzbek and Azeri etc.). All Turkic languages belong to the Ural-Altaic language family and are characteristically agglutinative languages which have productive inflectional and derivational morphology.

Most research about natural language processing and machine translation of Turkic languages focus on Turkish language. Mainly due to the fact that there is active ongoing research on the subject in Turkey, and they continue to produce valuable results. To date, machine translation systems implemented between Turkic languages has been scant, such as: Turkish to Azeri, Turkish to Crimean Tatar, Turkmen language to Turkish etc. Unfortunately, little computational research about Uygur languages exists. Turkic languages tend to have similar morphological structure and share some common word roots. The main shared properties include similar word order and syntactic structure. However, distinctions exist which prevent mutual intelligibility between these languages.

In order to implement this translation system, we utilized a frame-work which is favored for translation between closely related agglutinative languages. Thus, we implemented a morphological analyzer for Uyghur language with XEROX's Finite State Transducers (FST) tools. In this morphological analyzer we considered general cases for Uyghur languages and tagged Uyghur words with the same tags that were used for tagging other Turkic languages words. Thus, it will be easy to integrate this system to other Turkic languages. In order to improve the system's performance, we implemented a rule based morphological disambiguator, additionally, a disambiguator for word senses.

We have evaluated our system's performance using BLEU scores for 240 differently structured sentences.

As a result, a system has been determined which may successfully translate intermediate level Uyghur language into Turkish.

Keywords: *Machine translation, Turkic languages, Uyghur language, Turkish.*

Giriş

Dil insanlar arasında iletişimin sağlanmasındaki en önemli araçtır. Farklı dillerde konuşan insanların birbirlerine amaç, istek ve düşüncelerini anlatabilmeleri için her iki dili bilen çevirmen gerekmektedir. Bilgisayar teknolojisinin gelişmesiyle farklı diller arasında makineli çeviri yapan sistemlerin geliştirilmesi düşünülmüştür ve bu dala Bilgisayarlı Çeviri (BÇ) adı verilmiştir. İlk BÇ 1950 yılında Rusça ile İngilizce arasında geliştirilmiş ve 60 tümce için tüm anlamıyla doğru sonuç vermiştir (Hutchins, 2004). Bu çeviri sisteminin başarısı çok yüksek olmuştur ve sistemin mimarları 3–5 sene içerisinde yetkin bir çeviri sisteminin tamamlanabileceğini iddia etmiştir. Ancak gerçek proje çalışmasında, proje tahmin edildiği kadar kolay olmamıştır ve ünlü ALPAC raporu yayınlamıştır (Hutchins, 1995). Bu rapor BÇ işleminin çok yavaş, doğruluk oranın düşük ve maliyetinin insan tarafından yapılan çeviri maliyetinden daha yüksek olduğu açıklamıştır. Bu rapor BÇ araştırmaları üzerinde olumsuz bir etki yaratmış ve yaklaşık 10 yıl konu unutulmuştur.

Bilgisayar teknolojisinin gelişmesi ile yüksek hızlı ve yüksek bellek sığalı bilgisayarlar üretilmiştir. Bundan dolaylı Doğal Dil İşleme çalışmaları ve BÇ için önemli engeller aşılmıştır. BÇ ve DDİ çalışmaları bilgisayar bilimlerinin önemli bir dalı haline gelmiştir. BÇ için çok önemli sistemler geliştirilmiş olmakla beraber geliştirilen tüm sistemler sadece belli bir alanda iyi sonuç vermektedir. Örneğin İngilizce Fransızca dilleri arasında geliştirilen METEO sistemi hava tahmini ile ilgili bilgileri çevirmek için geliştirilmiştir ve çok iyi sonuç vermektedir (Chandioux, 1976). Ancak genel amaçlı bir BÇ sistemi henüz geliştirilememiştir. Şimdiye kadar geliştirilen sistemlerde, birbirine yakın diller arasında geliştirilen BÇ sistemlerinin başarısı daha yüksek olmaktadır ve sistemlerin geliştirilmesi de kolay olmaktadır.

Türk dilleri arasında BÇ

Türk dil ailesi Türkçe, Türkmence, Azerice, Özbekçe, Uygurca, Kırgızca ve Kazakça gibi dilleri içermektedir. Şimdiye kadar Türk dil ailesi için yapılan DDİ çalışmaları ve BÇ ile ilgili yapılan araştırmalar Türkçe[1] üzerinde yoğunlaşmıştır. Örneğin Türkçeden Azericeye (Hamzaoğlu, 1993), Türkçeden Kırım Tatarcısına (Altıntaş, 2000) ve en son olarak Türkmenceden Türkçeye (Tantuğ vd., 2008) BÇ sistemi geliştirilmiştir. Hamzaoğlu (1993) ve Altıntaş (2000) tarafından geliştirilen sistemler sözcük bazında aktarma yapmaktadır ve kural tabanlı çalışmaktadır. Türkçeden Tatarcaya geliştirilen BÇ sisteminde biçimbilimsel belirsizlik giderme çalışması yapılmadığından dolayı, sistem birden fazla çözüm üretmektedir. Tantuğ (2008) geliştirdiği çeviri sisteminde, kural tabanlı ve istatistiksel yöntemleri birleştiren Karma[2] yöntem kullanmıştır. Bu yöntemin çalışma adımlar kısaca aşağıdaki gibi özetlenebilinir.

1. Kaynak tümcede karakter düzeltme ve ayrıştırma işlemi yapılır.
2. Kaynak dildeki sözcüklerin biçimbilimsel çözümlenmesi yapılır ve çoklu sözcük grupları belirlenir.
3. Kaynak dilin biçimbilimsel çözümlemeleri hedef dile aktarılmak üzere kurallar tanımlanır.
4. Kök aktarma sözcüğünden yararlanarak, kaynak dilin kök sözcükleri, hedef dile çevrilir.
5. Türkçe derleme göre çalışan İstatistiksel Dil Modellerinden (İDM) yararlanarak, hedef dilde tümce oluşturmak için olasılığı en yüksek olan çözümleme seçilir.
6. Yapısal düzeyde hedef dil için tümce kuralları tanımlanır.
7. Türkçenin biçimbilimsel çözümleyicisinden yararlanarak görünen biçimdeki yüzeysel sözcükler üretilir.
8. Yüzeysel düzeyde çalışan tümce kuralları tanımlanır.

Bu yöntemin başarımın yüksek olmasının temel nedeni ise, Türkmencenin biçimbilimsel çözümleyicisi, Tantuğ ve diğerleri (2006), ile çözüm-

[1] Türkiye Türkçesinden bahis edilmektedir.

[2] Bu yöntem Tantuğ'un Doktora çalışmasında önerilmiştir ve asıl adı: Akraba ve Bitişken Diller Arasında Bilgisayarlı Çeviri İçin Karma Bir Model. Atıfta bulunurken kısaca "Karma" yöntem diye kullanılmıştır (Tantuğ, 2007).

lenen tüm bilgiler Türkçe derlem ile çalışan İDM ile hesaplanmaktadır ve en yüksek olasılığı olan çözüm seçilmektedir. Bu nedenle, biçimbilimsel ve anlamsal belirsizlik aynı anda giderilmektedir. Ayrıca, yapısal düzeyde tanımlanamayan kurallar, görünen düzeyde tanımlanmıştır. Sistemde kullanılan İDM modeli, hedef dil için tanımlanan bir derlem üzerinde çalıştığından ve bu adımdan sonraki çalışmalar, kaynak dilden tamamen bağımsız yapılabildiğinden, tüm bitişken diller arasında çeviri yapabilecek bir alt yapı sunmuştur. Örneğin, Türkmenceden Türkçeye çeviri için geliştirilen sistemde, Türkçe için yapılan çalışmalar üzerinde hiçbir değişiklik yapmadan, bir başka dilden Türkçeye çeviri sistemi geliştirilebilmektedir. Yapılması gereken tek çalışma ise, kaynak dil ile ilgili biçimbilimsel çözümleyicinin geliştirilmesi, aktarma kuralları ve kök aktarma sözcüğünün tasarlanmasıdır.

Bu çalışmada Uygurcadan Türkçeye bir BÇ sistemi geliştirilmiştir. Uygurca ile ilgili doğal dil çalışmalarına ilişkin herhangi bir çalışma bulunmamaktadır. Bundan dolayı, bu çalışma ile geliştirilecek çeviri sistemi için daha önceden yapılmış olan çalışmalardan özellikle Türkiye Türkçesi için yapılan çalışmalardan yararlanılmıştır. Türk dilleri her ne kadar birbirine benzese de aralarındaki farklar azımsamayacak kadar fazladır. Bundan dolayı bir dil için geliştirilmiş çeviri sistemi doğrudan bir başka dil için kullanılamamaktadır. Bu nedenle Uygurcanın dil yapısı araştırılmış ve Türkçe ile karşılaştırılmıştır (Orhun vd., 2009a). Uygur dili için en önemli etkenlerden biri de abecedir. Uygurca için birden fazla abece kullanılmaktadır. Okullarda ve resmi kurumlarında Arap abecesi kullanılırken, İnternet ve bilgisayar ortamlarında Latin abecesi ağırlıklı olarak kullanılmaktadır (Duval vd., 2006). Bu çalışmada tüm kurumlar tarafından ortak kullanılan Latin abecesi, yani "Uygur bilgisayar abecesi" kullanılmış ve bu abece aşağıda kısaca anlatılmıştır.

Uygur bilgisayar abecesi

Latin abecesinden Arap abecesine geçiş yakın tarihte gerçekleştirildiğinden (Kaşgarlı, 1992), Latin abecesi resmi olmasa da kullanılmaktadır. Özellikle İnternet teknolojisinin yaygınlaşmasıyla, İnternet ortamında yazılan Uygurca metinlerin çoğu Latin abecesi ile yazılmaya başlanmıştır. Bazı İnternet siteleri Uygur Arap abecesi ve Latin abecesi kullanmak üzere farklı iki abecede yayımlanmaya başlanmıştır. Bundan dolayı İnternet ortamında kullanılan Latin abecesi Uygur Kompuyotor İlmiy Jemiyeti (UKIJ) tarafından geliştirtmiştir (UKIJ, 2000).

Uygurcanın biçimbilimsel çözümleyicisi için XEROX sonlu durum makinesi kullanıldığından ve XEROX'un ASCII karakterlerini kullanmasından dolayı, Uygur bilgisayar abecesinde kullanılan çift karakterler tek karakterler ile temsil edilmiştir. Yani, ch →c ile, sh→S ile, ng→N ile, gh→G ile, zh→Z ile, ö→O ile, ü→U ile, é →E ile temsil edilmiştir.

Böylece oluşturduğumuz sistemde kullanılan Uygur abecesi büyük ve küçük karakterlerden oluşmakta ancak tek karakterli yapıdadır. Büyük ve küçük karakterler bir birinden farklı anlam taşımaktadır. Bundan dolayı farklı okunmaktadır. Bu karakterler sadece Uygurcanın biçimbilimsel çalışmalarında kullanıldığından, genelde çeviri yapıldıktan sonra ya da önceki metinlerde geçen karakterler ile değiştirilmektedir. Uygurca toplam 32 harfinden oluşmaktadır:

a e b p t j c x d r z Z s S G f q k g N l m n h o O u U w E i y

Uygur harfleri genel olarak, ünlü ve ünsüz diye iki ana kümeye ayrılırlar. Uygurcada 24 adet ünsüz harf vardır:

b p t j c x d r z Z s S G f q k g N l m n h w y

Ünsüz harfler kendi içersinde ünlü ünsüz harfler ve ünsüz ünsüz harfler diye iki türe ayrılırlar:
Ünlü ünsüz harfler:

b j d r z Z G g N l m n h w y

Ünsüz ünsüz harfler: p t c x s S f q k
Uygurcada sekiz adet ünlü harf vardır:

a e E i o O u U

Ünlü harfler seslendirilirken, dil durumunun değişmesine göre üç türe ayrılırlar:

Dil altı ünlü harfler : e O U
Dil arkası ünlü harfler : a o u
Dil arası ünlü harfler : E i

Ünlü harfler telaffuz edilirken, dudak şeklinin değişmesine göre iki türe ayrılırlar:

Yuvarlak dudaklı ünlü harfler : o O u U
Yuvarlak dudaklı olmayan ünlü harfler: a e E i

Uygurcada ekler eklenirken, ünlü harfler çok etkin olurlar. Uygurcada bazı ünlü harflerin kesin bir şekli yoktur. Türkçede ünlü harflerin sadece kalın ve ince diye iki şekli varken, Uygurcada üç farklı şekilleri bulunmaktadır. Örneğin, dil arası ünlü harfler kümesinde "E" ve "i" gibi iki harf bulunmaktadır, Bu harfler sözcüklerde geçen bazı ünsüz harflere göre değişmektedir. Yani beraber bulunduklarında ünsüz harflere göre bazen dil altı ünlü harf olarak kabul edilirken, bazı durumlarda ise dil arkası ünlü harf olarak geçmekte ve ekler dil arkası türüne eklenmektedir. Bu neden ile ünsüz harfler de birkaç farklı alt kümelere ayrılırlar (Tömür, 2003). Uygurca içerdiği sözcük yönünden başka Türk dillerine göre, özellikle Türkiye Türkçesine göre daha karmaşıktır. Uygurcaya, Arapça, Farsça, Rusça ve son zamanlarda ise, Çinceden pek çok sözcük girmiştir. Bundan dolayı, genel olarak, Uygurca sözcükler için geçerli olan kurallar uygulansa bile birçok sözcük kural dışıdır. Uygurca bitişken dildir ve eklerin eklenmesi belli kurallara göre gerçekleştirilir. Arapça, Farsça sözcüklerde ön eklerde bulunmaktadır. Bu çeşit sözcükler Uygurcada kayda değer bir yoğunluktadır. Bunların dışında Çinceden ve Rusçadan alınan sözcüklere kural tanımlama olanağı yoktur. Dolayısıyla her sözcük için özel kural tanımlamak gerekmektedir.

Uygurcanın biçimbilimsel çözümleyicisi ve iki düzeyli kurallar

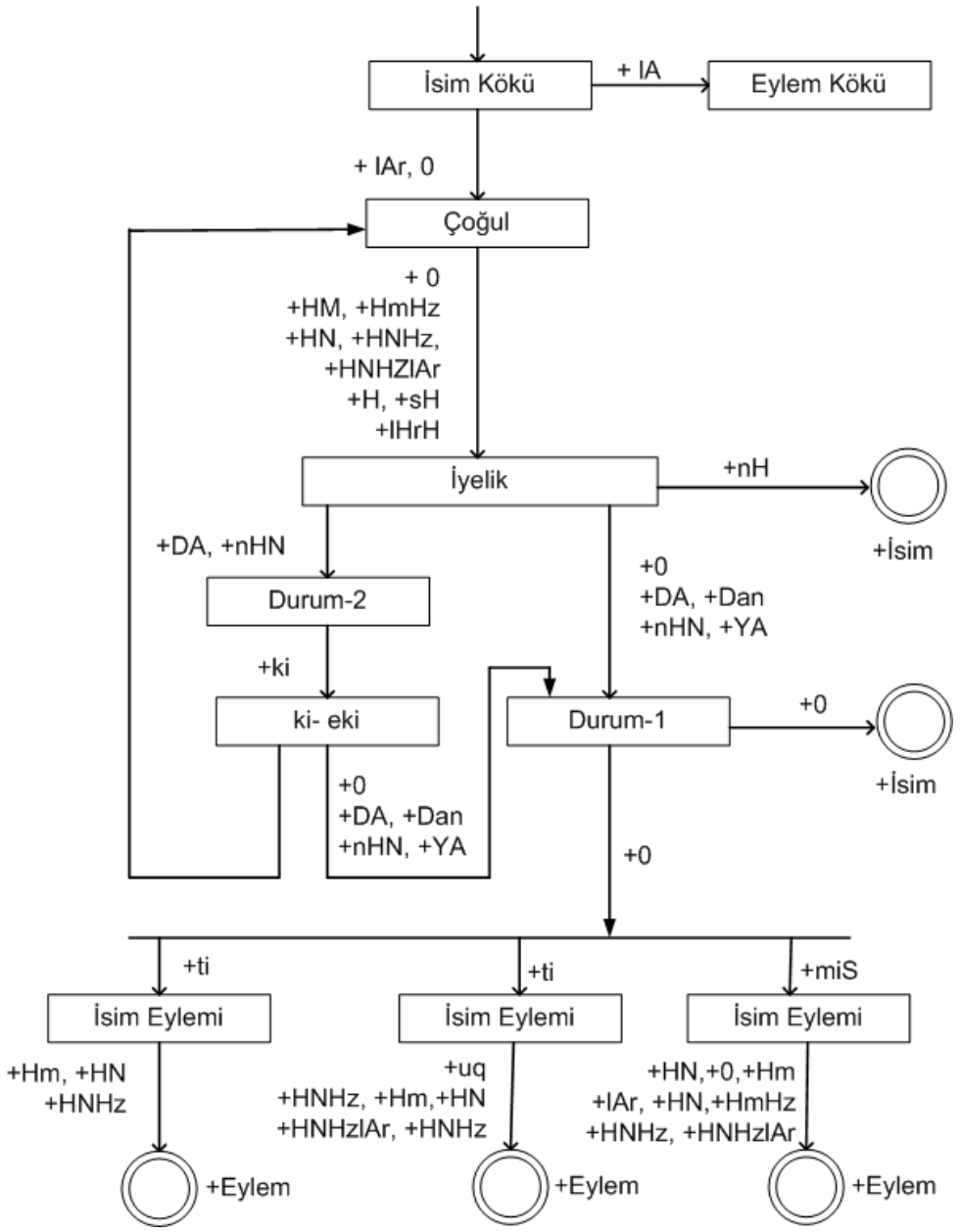
Uygurcanın biçimbilimsel çözümleyicinin tasarlanması için iki düzeyli biçimbilimsel çözümleme yöntemleri kullanılmıştır (Koskenniemi, 1983). Öncelikle Uygur dili dilbilgisine uygun olarak iki düzeyli kurallar tanımlanmıştır. Uygurca isimlerin biçimbilimsel çözümlenmesi Şekil 1'de verilmiştir.

Sekil 1'de gösterilen çözümleyicide, kareler bir ara durumu göstermektedir, yani son durumlar değildir. Bu durumlardayken geçerli sözcük olarak kabul edilemez. Çift yuvarlak çizgiler ise son durumlardır. Başlangıç durumdan, ekler eklenerek ya da ek almadan geçerek (yani boş geçerek "0") son duruma gelebiliyorsa, bu bir geçerli durum kabul edilebilir. Örneğin yapısal düzeye "kitapnH" sözcüğü gönderilirse, isimler söyle gerçekleşir: İlk önce çoğul eki alıp almadığına bakılır. Bu sözcükte çoğul eki "lAr" yok, dolayısıyla bir sonraki duruma "0" eki, yani "boş" ek ile geçiyor. Bir sonraki adımda ise "nH" eki aranmaktadır. "nH" eki alırken, çift yuvarlak çizgiye, yani "son" duruma gelir. Bundan dolayı "kitapnH" sözcüğü geçerli bir sözcük olarak kabul edilir ve biçimbilimsel bilgilerini geçiş yaptığı durumlardaki eklere göre yorumlanır. Bir sonraki adımda ise "nH" eki aranmaktadır. "nH" eki alırken, çift yuvarlak çizgiye, yani "son" duruma gelir. Bundan dolayı "kitapnH" sözcüğünü geçerli bir sözcük olarak kabul edilir ve biçimbilimsel bilgilerini geçiş yaptığı durumlardaki eklere göre yorumlanır. Şekil 1'de gösterilen "A, H, D,Y" harfleri ise yapısal düzeyde tanımlanan harflerdir. Bu harfler kural tanımlamaya kolaylık sağlamak için tanımlanmıştır ve görünen düzeyde görünmeyecektir. Örneğin, yapısal düzeydeki "A" harfi, görünen düzeydeki "a" ve "e" harfine denk gelmektedir. Bundan dolayı yapısal düzeyde "a" ve "e" harfi ise "A" ile temsil edilmektedir. Örneğin:

Yapısal düzey:	kitap+lAr
Görünen düzey:	kitaplar
Yapısal düzey:	qelem+lAr
Görünen düzey:	qelemler

Dolaysıyla yapısal düzeyden görünen düzeye geçerken "A" nın "a" ya da "e" olmasını dolaysıyla yapısal düzeyden görünen düzeye geçerken "A" nın "a" ya da "e" olmasını belirlemek için kural yazılmıştır. Sözcüklerin son hecesinde BACKV (Dil arkası ünlü harf) içerirse, "A" ise "a " olarak çözümlenir.

A:a=> [[:CONS]*[:BACKV][CONS]*]+(%+:0) [:CONS |CONS:]* _ ;



Şekil 1. Uygurca isimlerin sonlu durumlu makineler ile çözümlenmesi

Eğer sözcüklerin son hecesinde FRONTV (Dil arkası ünlü harf) içerirse, "A" harfi ise "e " olarak çözümlenir.

```
A:e=>[ [:CONS]*[:FRONTV] [CONS]*] +
        (%+:0) [:CONS | CONS:]* _ ;
```

İki düzey biçimbilimsel çözümleyici için tanımlanan kurallar paralel çalışmaktadır. Bundan dolayı bir dil için gereken tüm durumlar için kurallar yazılabilir. Uygurcada özellikle eylemler değişkendir. Bu nedenle genel durumlar dışında, özel durumların da göz önüne alınması zorunludur (Eziz, 2007; Belikiz, 2007). Geliştirmiş olduğumuz sistemde Uygurcanın genel durumu göz önüne alınmış ve Uygurcadan Türkçeye çeviri sistemi geliştirmek üzere biçimbilimsel çözümleyici ve iki düzeyli kurallar tanımlanmıştır (Orhun vd., 2009b; Orhun vd., 2009c).

Uygurcadan Türkçeye BÇ sistemi

Uygurcadan Türkçeye bilgisayarlı çeviri sistemi Şekil 2'de gösterildiği gibi Tantuğ'un (2007) önerdiği Karma yöntem üzerine belirsizlik gi-

derme yöntemleri eklenerek geliştirilmiştir. Bu nedenle, Türkçe için yapılan daha önceki çalışmalardan yararlanmış ve geliştirilecek çeviri sistemi ile ilgili çalışma Uygurca üzerinde yoğunlaşmıştır. Sistem ilk önce çeviri yapılacak tümcedeki sözcükler üzerinde karakter düzeltme işlemi gerçekleştirir. Örneğin,

u taghda at mindi.
u taGda at mindi.

Tümcede bulunan "taghda" sözcüğündeki "gh" çift karakteri biçimbilimsel çözümleyicinin çözümleyebileceği tek karakter "G" ile değiştirilir. Karakter düzeltme işleminden sonra, sözcük ayrıştırma işlemi gerçekleştirilir. Çünkü biçimbilimsel çözümleyici bir seferde sadece bir sözcük üzerinde biçimbilimsel çözümleme yapabilir. Yani biçimbilimsel çözümleyici çalışırken, "men", "taGda", "at", " mindi", "." gibi beş adet ayrı sözcük ve ayraç üzerinde çözümleme yapacaktır ve aşağıdaki çözümlemeleri üretecektir.

u+Pron+Pers+A3sg+Pnon+Nom
taG+Noun+A3sg+Pnon+Loc
at+Noun+A3sg+Pnon+Nom
at+Verb+Pos+Imp+A2sg
min+Verb+Pos+Past+A3sg
.+Punc

Çözümlemede, "u" sözcüğü şöyle yorumlanır: "u" bir zarf (+Pron), kişisel zarf (+Pers), tekil (+A3sg), şahıs eki yok (+Pnon) ve durum eki (+Nom) de yok. Başka sözcükler de çözümlendiği etiketlere göre nasıl özellik taşıdığı öğrenebilinir.

Biçimbilimsel belirsizlik giderme

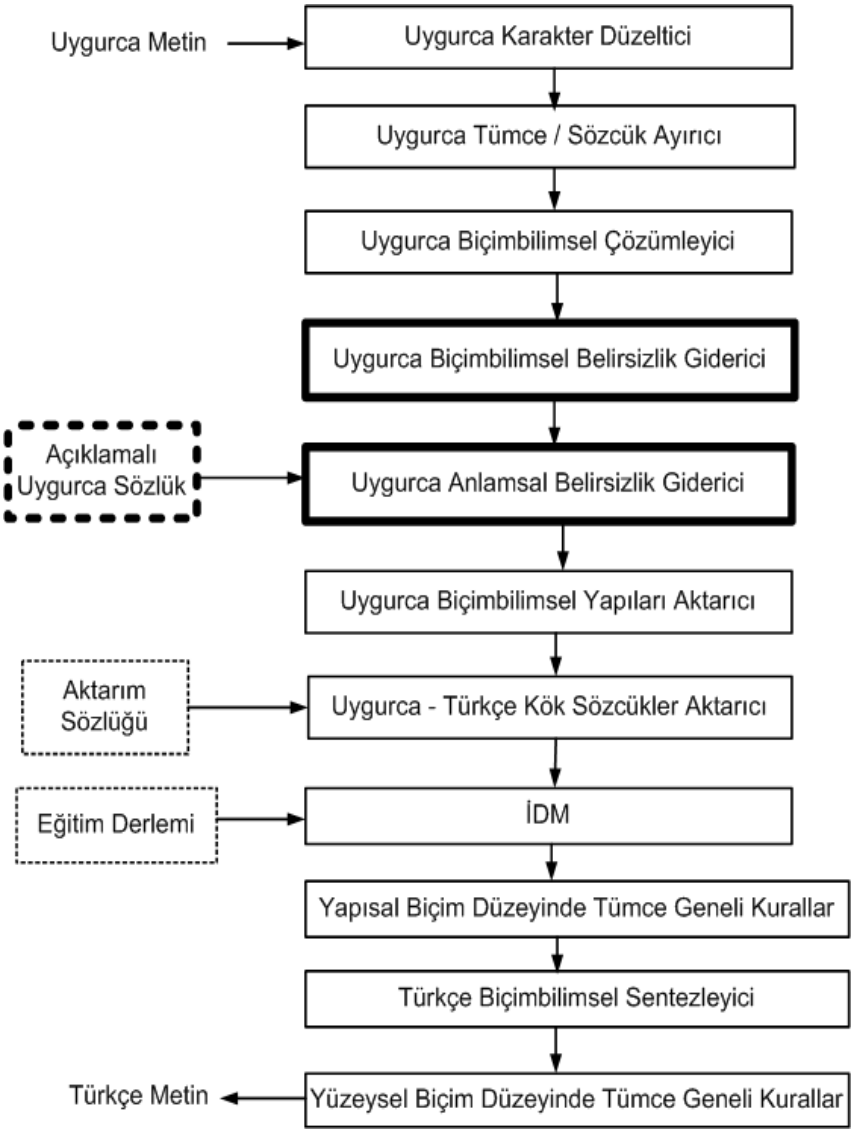
Biçimbilimsel çözümleyiciler genelde bir sözcük ile ilgili tüm bilgileri çözümlemelidir. Bu nedenle bir sözcük çoğu zaman birden fazla çözüm üretir. Bir sözcüğün birden fazla çözüm üretmesine, biçimbilimsel belirsizlik denir. Biçimbilimsel çözümlemeden bir sonraki çalışma bilgisayarlı çeviri sırasında, o sözcük ile ilgili biçimbilimsel çözümleme sırasında üretilenlerden birinin seçilmesi gerekir. Uygurca için geliştirilen biçimbilimsel çözümleyici ile yapılan hesaplamada, Uygurca sözcüklerin belirsizlik oranı yüzde 1,33 olarak tespit edilmiştir (Orhun vd., 2010). Bu nedenle Uygurca üzerinde BÇ çalışması yaparken, biçimbilimsel belirsizlik giderme çalışmasının yapılması gerekmektedir. Örneğin, bir önceki bölümde verilen örnek ile ilgili çözümlemede "at" sözcüğü ile iki faklı çözümleme üretilmiştir.

at+Noun+A3sg+Pnon+Nom
at+Verb+Pos+Imp+A2sg

İlk çözümlemede "at" sözcüğü isim (+Noun) olarak çözümlenirken, bir sonraki çözümde eylem (+Verb) olarak çözümlenmiştir. "at" sözcüğü isim olarak Türkçeye çevrildiğinde, "at", "ad" gibi anlam verirken, eylem olarak çözümlenirken "fırlat" olarak çevrilir. Bu nedenle "at" sözcüğü ile ilgili doğru çözümlemenin seçilmesi gerekir. Genelde biçimbilimsel belirsizliklerin giderilmesi için, istatistiksel ve kural tabanlı yöntemler kullanılır. Uygurca için derlem bulunmadığından dolayı kural tabanlı belirsizlik giderme yöntemi kullanılmıştır ve Uygur diline göre kurallar tanımlanmıştır. Kurallar genel olarak, sözcük etiketlerine, sözcüklerin tümcede olduğu yerlerine, sözcük türlerine göre çalışmaktadır. Örneğin, "at" sözcüğü ile ilgili olarak, "eylem" ya da "isim" olarak çözümlenmesi gerekir. "At" sözcüğü eylem olarak çözümlendiğinde, eylemlerin emir (+Imp) kipinde çözümlenmiştir. Ancak Uygurcada eylemler emir kipinde geldiğinde her zaman tümce sonunda gelir. Tümcede ise, "at" sözcüğünden sonra "bin" sözcüğü var. Yani "at" sözcüğü tümce sonunda değil. Bu nenenle "at" sözcüğünün eylem olarak çözümlenmesinin bu tümce için bir anlamı yoktur. Başka sözcükler ile ilgili sadece bir çözüm olduğunda, doğrudan geçerli çözüm olarak kabul edilir.

u+Pron+Pers+A3sg+Pnon+Nom
taG+Noun+A3sg+Pnon+Loc
at+Noun+A3sg+Pnon+Nom
min+Verb+Pos+Past+A3sg
.+Punc

Bu çalışmada tanımlanan kurallar ile Uygurcanın biçim bilimsel belirsizliğin giderim başarımı yüzde 96.86 olarak gerçekleştirilmiştir (Orhun vd., 2010).



Şekil 2.Uygurcadan Türkçeye çeviri sisteminin yapısı (Tantuğ, 2007)'den uyarlanmıştır

Sözcük anlamsal belirsizliğin giderilmesi

Bir dildeki bir sözcük bir başka dile çevrilirken birden fazla anlam içerebilir. Örneğin, Uygurcada bulunan "at" sözcüğü Türkçeye çevrilirken, hayvan anlamına gelen "at" ve insan ismi anlamına gelen "ad" olarak çevrilebilir. Bu çalışmada anlam belirsizliğinin giderilmesi için Lesk algoritması kullanılmıştır (Lesk, 1986). Lesk algoritmasının çalışma yöntemi, bir tümcede geçen tüm sözcükler, anlam belirsizliği giderilecek sözcüğün, örneğin "at" sözcüğünün, farklı anlamlarının tanımı için kullanılan sözcükler ile karşılaştırılır. Hangi anlam tanımında kullanılan sözcüklerle uyuşma sayısı fazla ise, o anlam seçilir. Uygurca için geliştirilen anlamsal belirsizlik giderici Şekil 3'de verilmiştir. Öncelikle Uygurca için bilgisayarla okunabilen sözlük (BOS) oluşturulmuş ve en çok kullanılan anlamları

Şekil 3. Anlamsal belirsizlik giderici

tanımlanmıştır. Örneğin, "at" sözcüğünün üç farklı anlamı aşağıdaki gibi tanımlanmıştır.

#at(0): adem we nerilerge birilgen isim. u balaGa at qoydi. atamaq. buni dep ataymiz.
#at(1): at azsa tayGa egiSir. at mingen.sUt emgici haywan. men at mindim. bu at tiz yUgUreydu.
#at(2): bir nesini atmaq, cOrimek, taSlamaq. miltiq atti. oq atti. taS atti. top atti, qaS atti.

Algoritma çalışırken, tümcede geçen, "u", "taGda" ve "mindi" sözcükleri, "at" sözcüğünün "at(0)", "at(1)" ve "at(3)" anlamlarının tanımlamak için kullanılan sözcükler ile karşılaştırılır. Karşılaştırma sonucu, "at(0)" anlamı için uyuşan sözcük sayısı 1, at(1) için 3 ve at(2) anlamı için 3 uyuşan sözcük bulunmuştur. Bu neden ile en yüksek uyuşan sözcük içeren anlam "at(1)" anlamı seçilmiştir.

Burada geçen "at(0)", "at(1)" ve "at(2)" sözcükleri sadece sözcük aktarmada kullanılan yapay sözcüklerdir. Bu yapay sözcüklerin Uygurca-Türkçe kök aktarma sözlüğünde aşağıdaki gibi tanımlanmıştır:

"ad" <- "at(0)"
"at" <- "at(1)"
"firlat" <- "at(2)"

Bu nedenle, Uygurca tümcede geçen "at" sözcüğünün Türkçe karşılığı olan, hayvan anlamına gelen "at" sözcüğü elde edilmiştir. Anlam belirsizlik giderme sonucu, Uygurca tümce ile ilgili biçimbilimsel çözümlemeler[3] aşağıdaki gibi oluşur.

u(0)+Pron+Pers+A3sg+Pnon+Nom
taG(0)+Noun+A3sg+Pnon+Loc
at(1)+Noun+A3sg+Pnon+Nom
min(0)+Verb+Pos+Past+A3sg
.+Punc

Lesk algoritmasında (asıl ve basitleştirilmiş) sözcük uyuşması hesaplanırken, tümcede geçen sözcükler ile sözcük tanımında geçen sözcükler doğrudan karşılaştırılır. Ancak bu yöntem Uygurca gibi bitişken diller için doğru sonuç vermemektedir. Bu nedenle bitişken diller için sözcük köklerine göre karşılaştırma yapan hesaplama yöntemi kullanılmıştır. Yapılan sınama sonuçlarında kök sözcüklere göre karşılaştırma yapan, işlevsel sözcük içermeyen ve basitleştirilmiş Lesk algoritmasının Uygurca için anlam belirsizliğini gidermede en iyi yöntem olduğu gözlemlenmiş ve yüzde 83 doğru sonuç vermiştir.

Biçimbilimsel yapı aktarıcı
Uygurca biçimbilimsel çözümleyici sözcükleri çözümlerken, Türkçenin biçimbilimsel çözümleyicisinde bulunmayan bazı etiketler kullanılmıştır. Çeviri sisteminin çalışması için, bu etiketlerin Türkçenin biçimbilimsel çözümleyicisi tarafından yorumlanması gerekir. Bu nedenle Uygurcadan Türkçeye biçimbilimsel yapı aktarma kuralları tanımlanmıştır. Örneğin, Uygurcada "siz" sözcüğü 2-şahsa karşı saygı anlamına gelir, tekildir ve aşağıdaki gibi çözümlenir:

siz+Pron+Pers+**A2sp**+Pnon+Nom

Ancak Türkçede "siz" ise saygı anlamı vermekle beraber çoğuldur.

siz+Pron+Pers+**A2pl**+Pnon+Nom

Bu neden ile biçimbilimsel aktarma kuralı ile Uygurcadaki "A2sp" etiketi "A2pl" ile değiştirilir.

Uygurca- Türkçe kök sözcük aktarıcı
Uygurca kök sözcükler Türkçeye aktarılırken, anlamsal belirsizlik giderme çalışmalarına göre aktarılmalıdır. Bu nedenle Uygurcadaki tüm kök sözcükler ve ek anlamlarına sayısal rakamlar ile ekleme yapılmıştır. Örneğin, bir önceki örnekte bulunan kök sözcüklerin aktarılması aşağıdaki gibi gerçekleştirilir:

o <-u(0)
dağ <-taG(0)
at <-at(1)
bin <- min
.+Punc <- .+Punc

Bu aktarma işlem sonucu, biçimbilimsel çözümlemeler aşağıdaki gibi olur.

[3] Tek anlamlı sözcükler için "0" anlamı seçilir varsayılan anlam olarak. İşlevsel sözcükler ve ayraçlar için belirsizlik giderme çalışması gerekmemektedir.

o+Pron+Pers+A3sg+Pnon+Nom
dağ+Noun+A3sg+Pnon+Loc
at+Noun+A3sg+Pnon+Nom
bin+Verb+Pos+Past+A3sg
.+Punc

İDM modeli

İDM modeli Türkçe eğitim derlemine göre çalışmaktadır ve en yüksek olasılığı veren biçimbilimsel çözümlemeyi seçmek için tasarlanmıştır. Dolaysıyla biçimbilimsel ve anlamsal belirsizlik giderme görevi görmektedir. Ancak geliştirilen bu çalışmada belirsizlik giderme çalışmaları kural tabanlı yöntemler ile giderildiğinden İDM için, eş anlamlı çözümlemeler dışında, sadece tek biçimbilimsel çözümlenmiş bilgi girişi yapılmaktadır. Bu nedenle İDM'in görevi sadece eş anlamlı sözcükler, örneğin, "insan" ile "adam", "yıl" ile "sene", "okul" ile "mektep" aralarında seçme işlevi görmektedir. Bu bakımdan İDM modeli geliştirilen sistemde pek etkin bir görev göstermemektedir. Ancak, Tantuğ'un (2007) geliştirdiği çerçeveye bağımlı kalmak için, geliştirilen model bu sisteme dahil edilmiştir.

Yapısal biçimde tümce kuralları

Kök aktarma ve İDM işlem sonucu elde edilen biçimbilimsel çözümlemeler doğrudan Uygurca tümceye göre yapıldığından, bu çözümlemeler üzerinde Türkçe tümce kurallarına göre ayarlama yapılması gerekmektedir. Uygurca ile Türkçenin söz dizimleri benzese de eklerin eklenme sırası farklı olmaktadır (Hengirmen, 2000; Öztürk, 1993; Tömür, 2003). Örneğin, Uygurcada "uniNdin sual soridim" tümcesi Türkçeye "ona soru sordum" olarak çevrilir. Yani Uygurcada bulunan "çıkma" durum eki "din", Türkçede "yönelme" durum eki "na" olarak çevrilmelidir. Bu kural tanımlanırken tümce bazında eklerin uyuşması aranır. Bu nedenle Uygurca eklerin Türkçeye doğru aktarılması için, Türkçe dil bilgilerine göre kurallar tanımlanmıştır (Orhun vd., 2009b).

Türkçe biçimbilimsel sentezleyici

Türkçe biçimbilimsel çözümlemelerden Türkçe sözcüklerin görünen biçimini oluşturmak için Oflazer'in geliştirdiği Türkçenin biçimbilimsel çözümleyicisi kullanılmıştır (Oflazer, 1995). Bu biçimbilimsel çözümleyici sonlu durumla makineler ile iki düzeyli kurallar kullanılarak geliştirildiğinden, ters yönde çalıştırıldığında, biçimbilimsel çözümlemelerden görünen biçimdeki sözcükler oluşturulmaktadır. Örneğin;

o+Pron+Pers+A3sg+Pnon+Nom
dağ+Noun+A3sg+Pnon+Loc
at+Noun+A3sg+Pnon+Nom
bin+Verb+Pos+Past+A3sg
.+Punc

Bu bilgilerden "o", "dağda", "at", "bindim", "." üretilir. Uygurca ile Türkçenin sözdizimi benzediğinden dolayı bu sözcüklerden doğrudan çeviri tümcesi elde edilir:

"ben dağda at bindim.".

Yüzeysel biçimde tümce kurallar

Yapısal düzeyde tanımlanamayan bazı kurallar yüzeysel biçimde tanımlanmıştır. Örneğin Uygurcada soru eki bir önceki sözcük ile bitişik yazılır. Ancak Türkçede ise bir önceki ekten ayrı yazılır ve ses uyumu sağlaması gerekmektedir. Ayrıca, Türkçede soru eki olarak dört faklı şekli (mı,mi,mu,mü) vardır. Bu tarz çözümlemelerin yapılması için yüzeysel biçimde tümce kurallarının tanımlanması gerekmektedir. Bu konu ile çalışma Türkçe için özel olduğundan ve daha önceden yapıldığından Tantuğ'un (2007) tanımladığı kurallar kullanılmıştır.

Çeviri örnekleri

Tasarlanan sistemin sınanması için değişik yapıda Uygurca tümceler ile sınanmış ve BLEU sonuçları elde edilmiştir. Örneğin:

UY: men bu mektepte 5 yil oqudum.
TR: Ben bu okulda 5 sene okudum.

Bu örnek ile ilgili olarak, çeviri başarımın hesaplamak için iki adet ölçün kullanılmıştır.

Ref1: ben bu okulda 5 sene okudum.
Ref2: ben bu okulda 5 yıl okudum.

Sistemin ürettiği aday çeviri bu ölçünler üzerinde BLEU sonucu hesaplandığında, başarımı

yüzde 100 olarak gerçekleştirilmiştir. Eş anlamlı sözcüler için aşağıda bulunan örnek verilmiştir.

UY: men bazardin 2000 dane yéngi kitap aldim.
TR: ben pazardan 2000 tane yeni kitap satınaldım.

Bu örnek ile ilgili olarak, çeviri başarımını hesaplamak için kullanılan referans tümceleri ise:

Ref1: ben pazardan 2000 tane yeni kitap aldım.
Ref2: ben pazardan 2000 adet yeni kitap satın aldım.

Sistemin ürettiği aday çeviri, bu ölçünler üzerinde BLEU sonucu hesaplandığında, başarımı yüzde 93.06 olarak gerçekleştirilmiştir.

Sonuçlar ve değerlendirmeler

Bu tez çalışmasında, ilk defa Uygurca için kapsamlı DDİ çalışması yapılmıştır. Uygurcanın biçimbilimsel çözümleyicisi geliştirilmiş ve bu biçimbilimsel çözümleyici kullanılarak Uygurcanın biçimbilimsel belirsizlik oranı somut olarak hesaplanmıştır (Orhun vd., 2009b; Orhun vd., 2009c). Belirsizlikleri gidermek için biçimbilimsel belirsizlik giderici geliştirilmiştir (Orhun vd., 2010). Uygurcada birden fazla anlam içeren sözcüklerin, kesin anlamını belirtmek için, anlamsal belirsizlik giderme çalışması yapılmıştır. Anlamsal belirsizlik gidermek için, bitişken diller için özel olarak kök sözcükler üzerinden hesaplama yöntemi geliştirilmiştir. Bu yöntemde kullanılmak üzere bir bilgisayarla okunabilen açıklamalı sözlük geliştirilmiştir. Geliştirilen anlamsal sözlük, 578 adet sözcük içermektedir.

Bu çalışma sonunda, Tantuğ'un (2007) önerdiği karma yöntem üzerine belirsizlik giderme çalışmaları eklenerek, karma yöntemler için kullanılan çerçeve kullanarak Uygurcadan Türkçeye çeviri yapabilen bir BÇ sistemi geliştirmiştir. Bu sistemin aktarma sözlüğünde toplam 1250 adet kök sözcük kullanılmıştır. Bu sözcükler eş anlamlı ve birden fazla anlam içeren sözcükleri de içermektedir.

Sonuç olarak bu çalışma ile Uygurca ile Türkçe arasında çeviri yapabilen bir BÇ sistemi geliştirilmiştir. Geliştirmiş sistem sadece Uygurcadan Türkçeye tek yönlü çeviri yapabilmektedir. Uygurca için DDİ ve BÇ araştırmalarında en temel sorun, şimdiye kadar yeterince çalışma ve alt yapının oluşturulmamasıdır. Uygurca derlem üzerinde çalışmalar başlanmış olsa bile genel kullanıma açık bir derlem bulunmamaktadır (Abaidula vd., 2005). Bu nedenle başka dillerden Uygurcaya BÇ yapabilmek için genel amaçlı ve herkese açık bir derlemin oluşturulması çok önemlidir.

Kaynaklar

Abaidula, Y., Rezwangul ve Sali, A., (2005). The Research and Development of Computer Aided Contemporary Uighur Language Tagging System. *Journal of Chinese Language and Computing* **15,** 4, 203-210.

Altıntaş, K., (2000). Turkish to Crimean Tatar Machine Translation System, *Yüksek Lisans Tezi*, Bilkent University, Ankara.

Belikiz, (2007). The 3253 different word forms Uygur Verb "qil", *Corpus Linguistics and Corpus Based Reseach*, Department of Linguistics, College of Anthropology, Xinjiang Normal University, Xinjiang, China.

Chandioux, J., (1976). Météo: Un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public, Meta, 21, 127-133.

Duval, J. R. ve Janbaz, W.A., (2006). An Introduction to Latin-Script Uyghur, *Middle East & Central Asia Politics, Economics, and Society Conference*, University of Utah, Salt Lake City, USA.

Eziz, G., (2007). Resistance to Borrowing of Uyghur Verbs, *Annual Conference*, University of Washington, USA.

Hamzaoğlu, İ., (1993). Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages, *Yüksek Lisans Tezi*, Bogazici University, İstanbul.

Hengirmen, M., (2000). *Türkçe Dilbilgisi*, Engin Yayıncılık, Ankara.

Hutchins, J., (2004). The first public demonstration of machine translation: *the Georgetown-IBM system, 7th January 1954. AMTA conference.*

Hutchins, J., (1995). Machine Translation: A Brief History, Concise history of the language sciences: from the Sumerians to the cognitivists. Edited

by E.F.K., Koerner ve R.E.Asher, Oxford, Pergamon Press.

Kaşgarlı, S.M., (1992). *Modern Uygur Türkçesi Grameri*, Orkun Yayınevi, İstanbul.

Koskenniemi, K., (1983). Two-Level Morphology : A General Computational Model for Word Form Recognition and Production, Department of General Linguistics, University of Helsinki.

Lesk, M., (1986). Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Code From an Ice Cream Cone, *Proceedings of the 5th Annual International Conference on Systems Documentation*, 24–26, ACM Press.

Oflazer, K., (1995). Two-level Description of Turkish Morphology, *Literary and Linguistic Computing*, **9,** 2, 137-148.

Orhun, M., Tantuğ, A.C, Adalı, E. ve Sönmez, A.C., (2009a). Computational comparison of the Uyghur and Turkish Grammar. *The 2nd IEEE International Conference on Computer Science and Information Technology*, **3**, 338-342, Beijing, China.

Orhun, M., Tantuğ, A.C., ve Adalı, E., (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, 811-816, Cairo, Egypt.

Orhun, M., Tantuğ, A.C. ve Adalı, E., (2009c). Rule Based Analysis of the Uyghur Nouns. *International Journal of Assian Language Processing* **19**, 1, 33-43.

Orhun M., Tantuğ, A.C. ve Adalı. E., (2010). Morphological Disambiguation Rules for Uyghur Language, *IEEE International Conference on Software Engineering and Service Science (ICSESS 2010)*, Bei Jing, China (Kabul edildi).

Öztürk, R., (1993). *Yeni Uygur Türkçesi Grameri*, Türk dil kurumu yayınları: **593**.

Tantuğ, A.C., (2007). Akraba ve Bitişken Diller Arasında Bilgisayarlı Çeviri İçin Karma Bir Model. Bilgisayar Mühendisliği Bölümü. *Doktora Tezi*. İstanbul Teknik Üniversitesi, İstanbul.

Tantuğ, A.C., Adalı, E., ve Oflazer, K., (2006). Computer Analysis of the Turkmen Language Morphology, FinTAL, *Lecture Notes in Computer Science*, **4139**, 186-193.

Tantuğ A.C., Adalı E., ve Oflazer K., (2008). Türkmenceden Türkçeye Bilgisayarlı Metin Çevirisi, *İTÜ Dergisi,***7,** 4, 83-94.

Tömür, H., (2003). *Modern Uygur Grammar (Morphology)*, Yildiz Teknik Üniversitesi, Fen-Ed Fak. T.D.E Bölümü, İstanbul.

UKIJ., (2000). http://ukij.org. (06.06.2000)