itüdergisi/***d***
mühendislik
Cilt:7, Sayı:4, 83-94
Ağustos 2008

Türkmenceden Türkçeye bilgisayarlı metin çevirisi

Ahmet Cüneyd TANTUĞ*[1], Eşref ADALI[1], Kemal OFLAZER[2]

[1] İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Programı, 34469, Ayazağa, İstanbul
[2] Sabancı Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi, 34956, Orhanlı, İstanbul

Özet

Diller arasında bilgisayar kullanılarak çeviri yapılması konusu, doğal dil işleme alanının en önemli dallarından bir tanesidir. Ancak teknolojideki ve yöntemlerdeki gelişmelere karşın, genel amaçlı, yüksek başarıma sahip çeviri sistemleri henüz genel kullanıma sunulamamıştır. Bunun temel nedeni, diller arasındaki büyük yapısal ve anlatım farklılıklardır. Bu noktadan hareketle, benzer diller arasında Bilgisayarlı Çeviri (BÇ) gerçeklemenin daha kolay olabileceği akla gelmektedir. Nitekim son yıllarda Çekçe-Slovakça, Çekçe-Lehçe, İspanyolca-Katalanca gibi çok yakın diller arasında yüksek başarımlı çıktılar üretebilen sistemler geliştirilebilmiştir. Üstelik bu sistemler, farklılıkların derin olduğu, Japonca-İngilizce gibi dil çiftleri arasında BÇ 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 Türkmence ile Türkçe dilleri arasında bir BÇ sistemi geliştirilmiştir. Söz konusu bu diller ne kadar benzer özellikler gösterse de, çözülmesi gereken farklılıklar azımsanmayacak boyuttadır. Türk Dilleri arasındaki farklılıklar, yukarıda anılan dil çiftlerinden daha fazladır ve karşılıklı anlaşılabilirlik söz konusu değildir. Sistem, hem kural tabanlı hem de istatistiksel bileşenlerden oluşan karma bir çeviri modeli kullanarak, Türkmence tümcedeki sözcüklerin sırasını değiştirmeden sözcük-sözcük Türkçeye aktarım yapılması ilkesini temel almıştır. Ancak bitişken yapılı Türk Dillerinin karmaşık biçimbilimsel özellikleri nedeniyle, diğer dillerde kullanılabilen basit doğrudan aktarım yöntemleri geliştirilerek kullanılmıştır. BLEU yöntemi ile sistemin başarım ölçümü yapılmış ve modelin başarılı sonuçlar üretebileceği gösterilmiştir.

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

* Yazışmaların yapılacağı yazar: Ahmet Cüneyd TANTUĞ. tantug@itu.edu.tr. Tel (212) 285 35 01
Bu makale, birinci yazar tarafından İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Programında tamamlanmış olan "Bitişken ve Akraba Diller Arasında Bilgisayarlı Çeviri İçin Karma Bir Model" adlı doktora tezinden hazırlanmıştır. Makale metni 09.05.2007 tarihinde dergiye ulaşmış, 20.06.2007 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 01.03.2009 tarihine kadar dergiye gönderilmelidir.

Machine translation from Turkmen language to Turkish

Extended abstract

Machine translation is a popular but hard field of natural language processing. Despite of the huge development of technology and inventions of new methods, general purpose full automatic Machine Translation (MT) systems do not exist. Today's MT systems either require post-editing or far from generating high quality translations, particularly of unrestricted texts. Primary reasons for that are the morphological, syntactical and lexical differences between different languages. The more distant language pairs are selected as source and target languages, the more complex methods or models must be used to build an MT system between those languages. Intuitively, this fact implies that MT between related languages can be easier than languages that have completely different structures (i.e. Japan and English).

Recently, MT between related languages like Czech to Slovak, Czech to Polish and Spanish to Catalan have been implemented and these studies showed that successful translations can be produced with relatively simpler efforts.

In this work our aim is building an MT model between Turkic languages. Some of the Turkic languages are Turkish, Azerbaijani, Uzbek, Turkmen, Kyrgyz, Kazakh and Uighur. All of the Turkic languages are agglutinative languages which have productive inflectional and derivational morphology. A high level of similarity can be observed between Turkic languages, especially in word order and syntactic structure. They have similar morphological structure and share some common word roots. However, some divergences preventing the mutual intelligibility are observed between these languages.

From the point of view of extending previous studies to Turkic languages, some serious problems emerge due to both agglutinative structure of the languages and resource scarcities. Except Turkish, most of the Turkic languages are computationally resource poor languages and that means the lack of training corpus, morphological analyzers, POS Taggers and machine readable dictionaries.

The model we have presented in this work is a hybrid model that has both rule based and statistical components. Since the word order of Turkic languages are almost same, a direct transfer approach is used in translation by means of word-by-word translation. Morphological processing, which can generate ambiguous results, is the first step of almost every NLP task for agglutinative languages. Then, the actual transfer is carried out by transferring the root words and morphological features to the target language. During the transfer of root word, another type of ambiguity, lexical ambiguity, is emerged in the process. As an exception of word-by-word transfer approach, some additional sentence level processing is done in order to translate Multi-Word Units (MWU) correctly. The two types of ambiguities are resolved by the disambiguation component that exploits Statistical Language Models (SLM) trained on the target language. In next component, a target language morphological generator produces the surface forms of the resulting candidate translation. As a last step, some work is done in sentence level because of some long distance dependencies and a number transfer rules for some phrase structures. The statistical disambiguation component is based on SLMs which are normally generated by using surface forms from the training corpus. But for an agglutinative language, such a training will heavily suffer from sparse data problem, so we propose some SLM types in which various parts of the full morphological parses are modeled. The performances of these types are investigated as well as the performance of the whole system.

To evaluate the practical performance of our model, we have implemented a generic MT framework based on our model and built a Turkmen to Turkish MT system by using this framework. The rule based modules of the system are implemented as Finite State Transducers (FST) using Xerox Finite State Toolkit. A Turkmen morphological analyzer is implemented in a two-level manner while an existing wide-coverage Turkish morphological analyzer is used in the generating direction as the target language morphological generator. We have used BLEU as the automatic evaluation metric for our MT system.

The results showed that general purpose MT between Turkic languages can be achieved, even relatively easier, and can generate high quality translations.

Keywords: *Machine Translation, Turkic languages, Turkmen Language, Turkish.*

Giriş

İnsanlar arasındaki iletişimi kolaylaştırmak adına, bilgisayarları kullanarak diller arası çeviri yapılması fikri, bilgisayarın kullanılmaya başlandığı ilk yıllardan bu yana birçok araştırmacının ilgisini çekmiştir. Bilgisayar yardımı ile doğal diller arasında çeviri yapılmasına Bilgisayarlı Çeviri (BÇ) adı verilmektedir. Ancak günümüz teknolojisi ve teknikleri ile bile yetkin bir BÇ sisteminin gerçeklenmesi çok uzaktır. Yetkin bir BÇ sisteminin temel özellikleri şunlardır:

- Otomatiklik: İnsan müdahalesine ya da düzeltmesine gerek kalmadan sonuç üretebilme özelliği
- Kaliteli Çeviri Yapabilme: Sistemin ürettiği çıktıların anlaşılabilir ve asıllarına uygun olma özelliği
- Geniş Kapsamlılık: Çeviri sistemi her türlü konuyu içeren genel metinler (makale, haber, hikâye, mektup vs.) üzerinde işlem görebilme özelliği

Diller arasında BÇ gerçeklemenin başlıca zorluğu, toplumlar arası kültür ve yaşayış farklılıklarının dillerine yansımış olmasıdır. Dili, bir anlaşma aracı olarak kullanan insanların hayat görüşleri, toplumsal yargıları, kültürleri, yaşayış şekilleri, yaşadıkları ortamların doğa koşulları gibi birçok farklı etken dillerini etkilemiştir. Bu etkiler, nesnelere, kavramlara ve eylemlere verilen isimlerin farklılaşmasına yol açtığı gibi tümce kuruluşlarını, vurguları ve anlatım biçimlerini de değiştirmiştir. Diller arası çeviri işlemi, ister bilgisayarla, ister insan emeği ile yapılsın, bir çok zorluklar içermektedir. Üstelik "doğru" çevirinin her zaman geçerli olan belirli bir nesnel ölçütü yoktur, aynı tümce farklı çevirmenler tarafından farklı şekillerde çevrilebilir.

Şimdiye dek yapılan çalışmalar sonucunda, eksik, hata oranı yüksek çıktılar üretebilen ya da kısıtlı kullanım alanına sahip uygulamalar gerçeklenmiş ve bu tür uygulamaların çeşitli alanlarda işe yaradığı görülmüştür. Gerçeklenen sistemlerin özellikleri incelendiğinde, yetkin çeviri sisteminin temel özelliklerinden bazılarından vazgeçilmesi yoluyla gerçeklenmiş sistemler üretildiği görülmektedir. Üretilen sistemler genel olarak üç ana amaçla kullanılmaktadır:

1) Yüzeysel Çeviri
   Bazı uygulamalarda, bir metnin yüzeysel bir çevirisi dahi iş görmektedir. Özellikle internet ortamından bilgi toplama sistemleri, kötü de olsa çeviri sonuçlarını kullanarak farklı dillerdeki içeriklere erişebilmeyi sağlamaktadır. Bu tür uygulamalarda yetkin sistemin "yüksek kaliteli çıktılar" üretebilme özelliği göz ardı edilmiştir.

2) Bilgisayar Destekli Dil Çevirisi
   Bilgisayarlı çeviri sistemlerinin bir diğer kullanım alanı da insan emeği ile yapılan klasik anlamdaki dil çevirilerini kolaylaştıran bir araç olarak kullanılmasıdır. Bu tür uygulamalarda bilgisayarlı çeviri sisteminin ürettiği sonuçlar, doğrudan kullanılmak yerine çevirmenler tarafından düzeltilerek kullanılır. Çevirmenlerin yapacakları bu değişiklikler, çoğu kez sıfırdan çeviri yapmaktan çok daha kolay olmaktadır. Yetkin çeviri sisteminin "otomatiklik" özelliğinden vazgeçilerek gerçeklenmiş bu sistemler, özellikle yüksek hacimli ve hızlı yapılması gereken çeviri işlerinde tercih edilir.

3) Kısıtlandırılmış Konularda Çeviri
   Çevrilecek metin türleri ve konuları kısıtlanarak gerçeklenen sistemlerde ise yetkin sistemin "geniş kapsamlılık" özelliği kullanılmamış olur. Bu tür sistemlerde konular ve hatta çevrilecek metinlerin dilbilgisi yapılarında bile bazı kısıtlamalara gidilir. Hava tahmini raporları, kullanılan dil açısından sabit kalıplardan ve hatta sabit sözcüklerden oluştuğundan bu kullanım için uygun bir örnektir.

Farklı dilbilgisel özellikler gösteren diller arasında (örneğin İngilizce-Japonca) BÇ gerçeklemesi bir çok zorluklar içerirken benzer diller (örneğin Çekçe-Slovakça) arasında BÇ daha kolay gerçeklenebilmektedir. Özetle, benzer diller söz konusu olduğunda, yetkin sistem özelliklerini taşıyan BÇ sistemlerinin geliştirilmesi da-

ha olanaklı görülmektedir. Ural-Altay dil ailesinde sınıflandırılan Türk dilleri de yapısal açıdan birçok benzerlik göstermektedir. Sözcük dağarcığı yönünden ortak kullanımların görüldüğü Türk Dilleri, bitişken yapıya sahip türetme ve çekim eklerine sahiptir. Sözdizimsel olarak incelendiğinde de özne-nesne-yüklem öğe sırasının korunduğu, hatta birçok durumda da sözcüklerin sıralarının değişmediği gözlenmektedir.

Dillerin sınıflandırılmasında kullanılan "karşılıklı anlaşılabilirlik" ilkesi uyarınca, örneğin Türkçe anlayabilen birisinin Türkmenceyi de anlayabilmesi gerekmektedir. Ancak birkaç istisna dışında (Türkçeyi bilen birisi Azericeyi kısmen anlayabilmektedir) bu önerme geçerli olmadığından, bütün bu benzerliklere rağmen Türk Dilleri ayrı birer dil olarak kabul görmüştür.

Dillerin yapısındaki bu benzerliklerden yararlanılarak Türk Dilleri arasında gerçeklenecek bir BÇ sisteminin başarımının yüksek olacağı tezinden yola çıkılmış ve Türkmenceden Türkçeye çeviri yapan bir BÇ sistemi gerçeklenmiştir.

Benzer diller arasında çeviri

Benzer diller arasında yapılan ilk bilgisayarlı çeviri çalışması, Çekçe ile Rusça arasında bilgisayarlı çeviri sistemi olan RUSLAN sistemidir (Hajič, 1987). Anaçatı işletim sistemleri ile ilgili her türlü belgenin Rusçaya çevrilmesi amacıyla gerçeklenen bu sistemde, kural tabanlı biçimbilimsel çözümleyici, çeviri ve biçimbilimsel üretici modülleri ile Çekçe sözdizimsel ayrıştırıcı ve Rusça birleştirici bileşenleri bulunmaktadır. Çalışmanın sonuçlarında, hedef dilde oluşturulan tümcelerin %40'ının doğru olarak çevrildiği, %40'ının bir kullanıcı tarafından düzeltilmesi gereken ufak hatalar içerdiği, geri kalan %20'lik bölümün ise tamamen baştan çevrilmesi gerektiği bildirilmiştir.

Bir diğer çalışmada ise birbilerine çok benzeyen iki dil olan Çekçe ve Slovakça arasında bir BÇ sistemi geliştirilmiştir (Hajič vd., 2000). ČESĬLKO adı verilen bu çalışmada, kaynak ve hedef dil özellikleri hemen hemen aynı olduğu için daha basit yöntemler kullanılmıştır. Bu yöntemlerin temelinde, sözcük bazında çeviri yapılmaktadır. Aşağıda bu sistemi oluşturan bileşenler tanıtılmıştır:

1) Çekçe biçimbilimsel çözümleme
2) Çekçe biçimbilimsel belirsizliğin giderilmesi
3) Konuya özel aktarım sözlüğü (tek ve birden fazla sözcükten oluşan girdiler)
4) Genel amaçlı aktarım sözlüğü
5) Slovakça biçimbilimsel üretim

Eşadlı sözcükler çevrildikten sonra da gene eşadlı kaldıkları için iki dil arası çeviride herhangi bir anlam belirsizliği oluşmadığı belirtilmiş ve aktarım sözlüğü, bire-bir aktarım yapacak şekilde geliştirilmiştir. ČESĬLKO sisteminin başarımı, Çekçe-Slovakça arasında %90 civarında rapor edilmiştir. Çalışma, daha sonra diğer Slav dilleri için de genişletilmiş, ilk olarak gene Çekçeye benzer bir dil olan ve Hint-Avrupa dil ailesinin Slav kolunda Çekçe ile aynı alt grupta sınıflandırılan Lehçe hedef dil olarak seçilmiştir. Bir sonraki adımda ise Çekçe ile diğerlerine göre daha az benzerlik gösteren ancak gene Hint-Avrupa dil ailesinin Baltık kolunda sınıflandırılan Litvanyaca hedef dil olarak kullanılmıştır (Hajič vd., 2003). Çalışma sonucunda Çekçe-Lehçe başarımı %71, Çekçe-Litvanyaca başarımı ise %69 olarak belirtilmiştir. Litvanyaca dilinin farklı yapısından dolayı, çeviri sistemine çok kapsamlı olmayan bir Çekçe sözdizimsel çözümleyici eklenmiştir.

Benzer diller arasında gerçeklenen bir diğer çalışma ise Hint-Avrupa dil ailesinin Romans kolunun İber-Romans alt grubunda beraber sınıflandırılan İspanyolca ile Katalanca dilleri arasında yapılmıştır (Canals-Marote vd., 2000). Sistem bileşenlerinden temel işlevlere sahip olan biçimbilimsel (Canals-Marote vd., 2000) çözümleyici, çeviri sözlüğü, biçimbilimsel üretici ve son işleme bileşenleri Sonlu Durumlu Makineler (SDM) kullanılarak gerçeklenmiştir. Gerçekleştirilen sistemin temel bileşenleri aşağıda listelenmiştir:

1) Biçim ayıklayıcı
2) Biçimbilimsel çözümleyici

3) Sözcük türü etiketleyici
4) Çeviri sözlüğü
5) Örüntü işleme bileşeni
6) Biçimbilimsel üretici
7) Son işlemler bileşeni
8) Biçim yapıştırıcı

Çeviri, temelde sözcük bazında yapılırken bazı durumlarda sözcük öbeklerinin çevirisi yapılmaktadır. Sözcük türü etiketleyici bileşeninde, 2-gram ve 3-gramlardan oluşturulmuş Hidden Markov Modeli (HMM) kullanılmıştır. Sözcüklerin çevrilmesi aşamasında kullanılan çift dilli aktarım sözlüğü, kaynak dildeki her bir sözcük için hedef dilde sadece bir sözcük karşılık düşürmektedir. Bunun nedeni olarak da, farklı anlamlara sahip eşadlı sözcüklerin her iki dilde de aynı şekilde ifade edilmesi gösterilmiştir. Bu projenin sonuçlarında İspanyolcadan Katalancaya çeviride (üretilen metnin kabul edilebilir olması için gereken sözcük ekleme, silme ya da değiştirme adetleri bazında ölçülen) hata oranı %5 iken ters yönde (Katalanca-İspanyolca) çevirilerde daha kötü bir başarı sağlandığı belirtilmiştir.

Yukarıda değinilen BÇ sistemlerinin tamamı, aktarım sözlüklerinde bire-bir karşılıklar kullanmaktadır. Yani, kaynak dildeki her bir sözcüğün hedef dilde sadece bir karşılığı bulunmaktadır. Seçilen dil çiftleri arasında bire-bir sözcük karşılığının olmasının çoğu durumda yeterli olduğu belirtilmiş ve bu yüzden anlamsal farklılıkların aktarılması konusunda herhangi bir işlem gerçeklenmemiştir. Ayrıca özellikle Çekçe ekseninde yapılan çalışmada kaynak dil olarak hep Çekçe kullanılmış, diğer dillerden Çekçeye çeviri konuları incelenmemiştir. Bunun nedeni belirtilmemiş olsa da, seçilen hedef diller için biçimbilimsel çözümleyici ve sözcük türü etiketleyici araçlarının mevcut olup olmaması, bu konuda önemli rol oynamaktadır.

Türk dilleri arasında BÇ çeviri konusunda yapılan ilk çalışma Türkçeden Azericeye olmuştur (Hamzaoğlu, 1993). Bir başka çalışma ise Kırım Tatarcası ve Türkçe arasındadır (Altıntaş, 2000). Bu çalışmalarda da sözcük bazında çeviri yöntemi gerçeklenmiştir. İkinci çalışmada sadece Türkçeden Tatarca'ya çeviri yapılmış, Tatarcadan Türkçeye çeviri gerçeklenmemiştir. Türkçe için Oflazer'in biçimbilimsel çözümleyicisi kullanılmış (Oflazer, 1995) ve Kırım Tatarcası için biçimbilimsel çözümleyici/üretici tasarlanmıştır (Altıntaş ve Çiçekli, 2001). Kullanılan biçimbilimsel çözümleyiciler kural tabanlıdır ve SDM temelli tasarlanmışlardır. Çeviri sözlüğünden, kaynak dilde eklerine ayrılmış sözcük kökünün karşılığı bulunmakta ve bu karşılık ile ilgili eklerin hedef dildeki karşılıkları birleştirilerek sözcük Tatarca olarak yeniden oluşturulmaktadır. Ancak bu çalışmada, biçimbilimsel belirsizliğin giderilmesi için herhangi bir yöntem kullanılmadığı için her sözcüğün olası her çözümlemesi Tatarcaya çevrilmektedir. Bu ise, bir giriş tümcesi için birden fazla karşılığın çıkması anlamına gelmektedir. Anılan çalışma da, yukarıda tanıtılan diğer sistemlerde olduğu gibi kaynak dildeki bir sözcüğe hedef dilde sadece bir karşılık eşleştirmektedir.

Türk dilleri

Türk dil ailesi, çoğunlukla Orta Asya coğrafyasına yayılmış ve toplamda yaklaşık 180 milyon insanın kullandığı, aynı temelleri paylaşan ve birçok benzer özelliği bulunan dillerden oluşmaktadır. Yaygın kullanılan bazı Türk dilleri Tablo 1'de verilmiştir.

Tablo 1. Başlıca Türk dilleri

Dil	Kullanıldığı Bölge	Kişi Sayısı
Türkçe	Türkiye	72 milyon
Azerice	Azerbaycan	24.3 milyon
Azerice	İran	7 milyon
Türkmence	Türkmenistan	6.4 milyon
Kazakça	Kazakistan	8 milyon
Kırgızca	Kırgızistan	2.6 milyon
Uygurca	Doğu Türkistan	7.6 milyon
Özbekçe	Özbekistan	18.5 milyon
Özbekçe	Afganistan	1.4 milyon
Çuvaşça	Rusya	2 milyon

Günümüz Türkiye Türkçesine en çok benzeyen diller, Türkçe ile aynı alt kol içerisinde yer alan Azerice, Türkmence ve Gagauz Türkçesidir. Özbekçenin günümüz Türkçesine yakınlık dere-

cesi ortadır. Kırgızca ve Kazakça, coğrafi olarak Türkiye'ye uzak oldukları için Türkçeye, Özbekçe ve Türkmenceden daha uzaktır. Kıpçak grubundaki diller arasında Türkçeye en yakın dil Kırım-Tatar Türkçesidir. Yakutça ve Çuvaşça ise Türkçeye en uzak dillerdir (Hengirmen, 2000).

Türkmenceden Türkçeye BÇ sistemi

Doğrudan aktarım temelinde gerçekleştirilen çeviri sisteminin temel bileşenleri Şekil 1'de verilmiştir. Bu sistemin bileşenleri ve görevleri aşağıda kısaca tanıtılmıştır.

Türkmence biçimbilimsel çözümleyici

Çeviri işleminin gerçeklenmesi için tümcedeki sözcüklerin biçimbilimsel açıdan çözümlenmesi, köklerinin ve eklerinin belirlenmesi gereklidir. Bu amaçla iki-düzeyli (two-level) yöntemlerle Sonlu Durumlu Dönüştürücü (SDD) tabanlı bir Türkmence biçimbilimsel çözümleyici geliştirilmiştir (Tantuğ vd., 2006a). Yüzeysel biçimde girişi olan sözcüğün aşağıdaki gibi yapısal biçimdeki karşılığı üretilmektedir:

```
Yüzeysel: eñrejekdirin
Yapısal : eñre+Verb+Pos+Fut+Cop+A1sg
```

Biçimbilimsel çözümleme sonucunda bir yüzeysel biçim için birden fazla yapısal biçim üretilebilir. Buna biçimbilimsel belirsizlik adı verilir.

Türkmence çoklu sözcük grubu işleyici

Sözcük bazında yapılan çeviri işlemleri sırasında birden fazla sözcükten oluşan yapıların hatalı aktarılmaları söz konusu olmaktadır. Örneğin aşağıdaki Türkmence iki sözcük, Çoklu Sözcük Grubu (ÇSG) olarak değerlendirilmeli ve Türkçeye buna uygun olarak çevrilmelidir:

```
gürüm-jürüm          (gizli)
gürüm-jürüm bolmak   (kaybolmak)
```

ÇSG'lerin belirlemesi yönünden en önemli zorluk, ÇSG'lerin çekime uğrayabilmesidir:

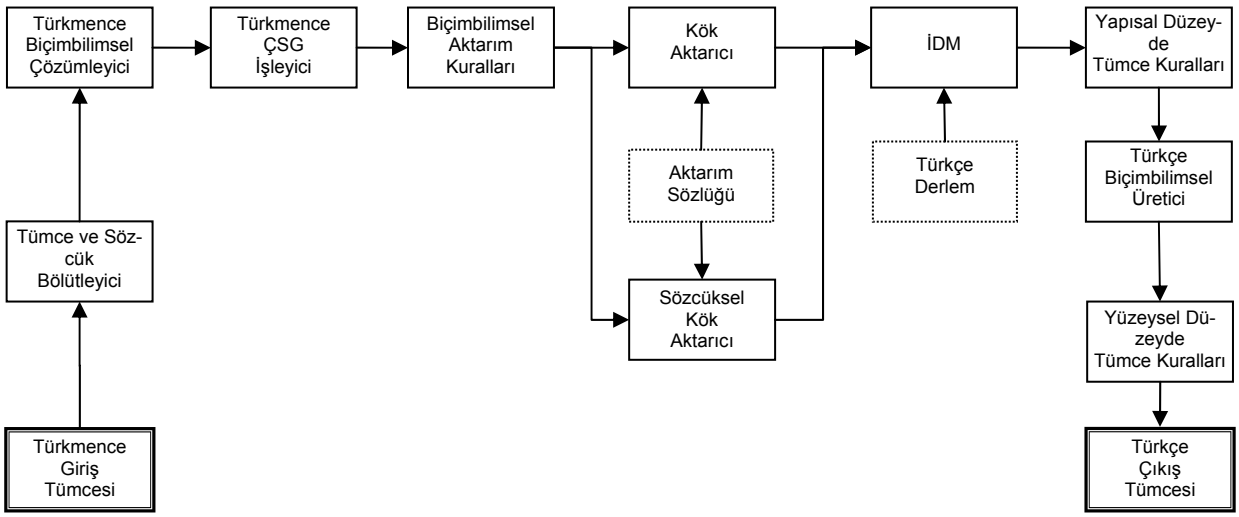
```
gürüm-jürüm bolmakçy (kaybolmayı düşünüyor)
gürüm-jürüm bolypdy  (kaybolmuştu)
```

ÇSG'lerin belirlenmesi için Türkçe üzerine yapılmış olan bir çalışma Türkmenceye uyarlanmıştır (Oflazer vd., 2004).

Biçimbilimsel aktarım kuralları

İki dil arasındaki biçimbilimsel farklılıkların giderilerek Türkmence sözcüklerin biçimbilimsel çözümlemelerinin Türkçe biçimbilimsel üreticinin beklediği biçime dönüştürülmesini sağlayan aktarım kuralları, SDD olarak tasarlanmıştır. Örneğin Türkmencede emir kipinin 1. tekil ve çoğul kişi çekimleri, Türkçe'deki istek kipine karşılık gelir:

```
"+Opt+A1sg" <- "+Imp+A1sg" .o.
"+Opt+A1pl" <- "+Imp+A1pl";
```



Şekil 1. Türkmenceden Türkçeye bilgisayarlı çeviri sisteminin temel bileşenleri

Bir diğer örnek ise, Türkmence'de "**+makçy/+mekçi**" ekleri ile kullanılan ancak Türkçe karşılığı "bir eylemi yapmayı düşünmek veya tasarlamak" olarak gösterilen eylem çekiminin aktarılmasıdır. Bu aktarım için aşağıdaki kural kullanılmaktadır:

```
"^DB+Noun+Inf1+A3sg+Pnon+Nom_iste+Verb+Pos+
Prog1+A3sg" <- "+Think+Anon";
```

Bu ve benzeri biçimbilimsel dönüşümleri yapmak üzere toplam 24 kural tanımlanmıştır.

Kök aktarıcı

Biçimbilimsel çözümlemesi yapılmış Türkmence sözcük köklerinin Türkçeye aktarılmasını sağlayan kurallar, SDD'ler ile gerçeklenmiştir. Örnek bir aktarım kuralı aşağıda verilmiştir:

```
"tatlı" <- "Yakymly"
```

Bu aktarım kurallarında sözcük türlerinin kullanılması, sözcüksel belirsizliği azaltmaktadır. Yazılan kurallar bu ilke çerçevesinde oluşturulmuş ve kuralların sağ bağlamları sözcük türleri ile kısıtlandırılmıştır:

```
"gri" <- "boz" \/  _ "+Adj"   .o.
"sil" <- "boz" \/  _ "+Verb"
```

Bu sayede sistemin rastladığı bütün "***boz***" köklerini, "***gri***" ve "***sil***" kökleri ile değiştirmesinin önüne geçilmiş ve aktarılacak sözcüğün sıfat ya da eylem olma durumuna göre sadece uygun karşılıkların dönüştürülmesi sağlanmıştır. Ancak gene de aktarım sözlüğünde, bir Türkmence sözcük için birden fazla Türkçe karşılık olabileceği için, kök aktarıcı bileşenlerin her ikisinde de sözcüksel belirsizlik ortaya çıkmaktadır.

Sözcüksel kök aktarıcı

Uygulamada ortaya çıkan bazı durumlar göstermiştir ki bir takım sözcükler için sadece sözcük kökünü değiştiren basit bir kural yeterli olmamaktadır. Örneğin Türkmencedeki ***ulumsy*** sözcüğü, Türkçedeki ***kibirli*** sözcüğünün karşılığıdır. Standart kurallar uygulanarak sadece sözcük kökü değiştirildiğinde aşağıdaki dönüştürme işlemi gerçeklenir:

```
kibirli+Adj <- ulumsy+Adj
```

İlk bakışta göze çarpan herhangi bir sorun olmamasına karşın, oluşan yapısal biçimdeki sözcük, Türkçe biçimbilimsel üretici tarafından yüzeysel biçime dönüştürüleceği zaman herhangi bir çıktı üretilememektedir. Bunun altında yatan neden ise, Türkçedeki ***kibirli*** sözcüğünün aslında türemiş bir sözcük olması ve bu sözcüğün doğru yapısal biçiminin aşağıdaki gibi olmasıdır:

```
kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With
```

Bu sorunun düzeltilmesi için, Türkmencedeki ***ulumsy*** sözcüğü için aşağıdaki gibi özel bir kural oluşturulmalıdır:

```
"kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With"<-
"ulumsy+Adj"
```

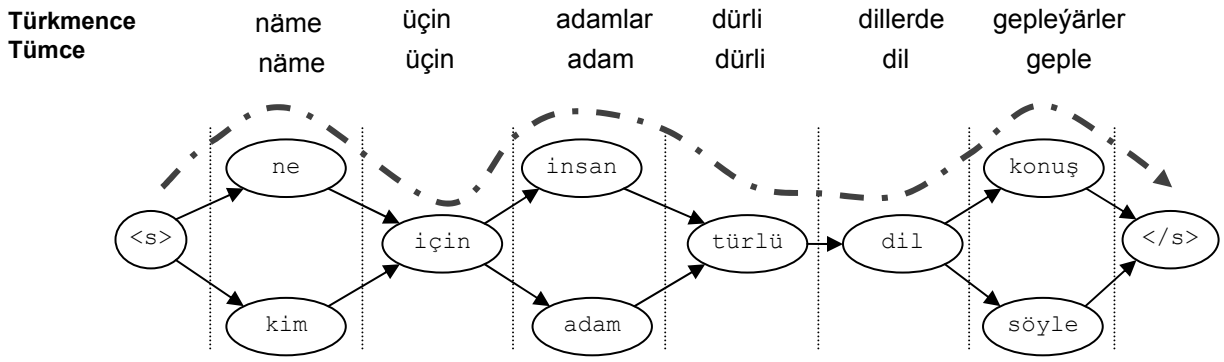
Örnekte açıklandığı gibi sözcüğe bağlı özel durumları kotaran kurallar, sözcüksel kurallar (lexicalized rules) olarak adlandırılmıştır. Ancak her iki dilde de ortak olan türetme ekleri ile türetilebilecek sözcükler için ayrı kuralların oluşturulmasına gerek yoktur. Örneğin Türkmencedeki ***+lyk*** eki ile Türkçedeki ***+lık*** eki, sıfattan isim yapan ve aynı anlama sahip iki yapım ekidir. Dolayısı ile Türkmencede bulunan ***ulumsylyk*** sözcüğünün karşılığı da ***kibirlilik*** sözcüğüdür. Her iki sözcüğün biçimbilimsel çözümlemesi aşağıda belirtilmiştir:

```
ulumsy+Adj
   ^DB+Noun+Ness+A3sg+Pnon+Nom
kibir+Noun+A3sg+Pnon+Nom^DB+Adj+With
   ^DB+Noun+Ness+A3sg+Pnon+Nom
```

Örnekten de görüldüğü gibi, kalın olarak gösterilmeyen biçimbilimsel yapılar aynıdır. Dolayısı ile bu iki sözcük için ayrı bir sözcüksel aktarım kuralı hazırlanmasına gerek yoktur, yukarıda anlatılan ve ***ulumsy*** sözcüğünü aktaran sözcüksel aktarım kuralının çalışması yeterli olmaktadır.

İstatiksel dil modeli

Çeviri sisteminin bileşenleri içerisinde, Türkmence biçimbilimsel çözümleyici ve kök aktarımı bileşenlerinin çıktıları belirsizlik içermektedir:



Şekil 2. En yüksek olasılıklı Türkçe tümcenin oluşturulması

Gerek biçimbilimsel belirsizlik, gerekse de sözcüksel belirsizliğin giderilmesini amaçlayan bu bileşen, istatistiksel yöntemlerle en olası sözcük dizisini (yani tümceyi) belirler. Bu amaçla, istatistiksel BÇ, ses tanıma gibi birçok alanda başarısı kanıtlanmış İstatistiksel Dil Modelleri (İDM) kullanılmıştır (Tantuğ vd., 2006b). Aktarım sistemindeki İDM bileşenine girdi olarak, kaynak dildeki tümcenin bütün sözcüklerinin aday çevirileri gelir. Bileşenin çıktısı olarak ise tüm kombinasyonlar içerisinden, seçilen İDM'ye göre en yüksek olasılığa sahip tümce üretilir. Olası tüm kombinasyonların tamamının olasılıklarının hesaplaması yerine, aday sözcüklerden bir Hidden Markov Modeli (HMM) oluşturularak üzerinde Viterbi algoritmasının çalıştırılmasıyla en yüksek olasılıklı sözcük dizisi elde edilebilir (Şekil 2). İngilizce ve Almanca gibi dillerden farklı olarak, Türkçe gibi bitişken diller için İDM oluşturulurken sözcüklerin yüzeysel biçimlerinin kullanılması seyrek veri sorununa yol açmaktadır. Bu yüzden eğitim verisi olarak sözcüklerin yüzeysel biçimleri yerine, sözcüklerin kökleri ve diğer bazı biçimbilimsel özelliklerin kullanılması yoluna gidilmiştir. Türkçede her sözcük, kök ve bir veya birden fazla çekim grubundan oluşmaktadır (Hakkani-Tür vd. 2002). Çekim grupları birbirlerinden ^*DB* ile ayrılmaktadır [1]:

$$\text{kök+ÇG}_1\text{\^{}DB+ÇG}_2\text{\^{}DB+…\^{}DB+ÇG}_n$$

Burada $ÇG_i$, sözcük türü ve çekim özelliklerini de içeren ilgili çekim grubunu göstermektedir. Örnek olarak, "***yararlanmanın***" gibi yüzeysel bir biçimin biçimbilimsel çözümleme sonucu üretilen yapısal biçimi aşağıda verilmiştir:

```
yarar+Noun+A3sg+Pnon+Nom                ÇG1
  ^DB+Verb+Acquire+Pos                  ÇG2
  ^DB+Noun+Inf2+A3sg+Pnon+Gen           ÇG3
```

Bu örnekte, isim türlü yarar sözcüğünün sözcük türü, ***+lan*** yapım eki ile önce eyleme daha sonra da ***+ma*** mastar eki ile de tekrar isme dönüşmüştür. Bu dönüşme süreci içerisinde oluşan her sözcük türünün de kendisine ilişkin çekim özellikleri bulunabilir. Türetilmiş bir sözcüğün etkin sözcük türü, son ÇG'nin sözcük türü olarak kullanılır (örneğin etkin sözcük türü "isim"dir). Bu yapıdan hareketle 5 farklı İDM türü hazırlanmıştır (Tablo 2).

Tablo 2. İDM türleri

İDM	Kullandığı Biçimbilimsel Yapılar
Tip I	Kök
Tip II	Son ÇG'deki sözcük türü
Tip III	Son ÇG
Tip IV	Kök ve Son ÇG'deki sözcük türü
Tip V	Kök ve Son ÇG
Tip VI	Kök hariç diğer tüm biçimbilimsel özellikler

Uygulamada farklı tiplerdeki dil modellerinin, Türkmenceden Türkçeye çevirideki belirsizlikleri giderme başarımları incelenmiş ve en yüksek başarımı sağlamak için bu dil modellerinin nasıl kullanılacağı araştırılmıştır.

[1] DB = türetme sınırı (derivation boundary)

Yapısal düzeyde çalışan tümce kuralları
Gerçeklenen BÇ sisteminde tümcenin sözdizimsel çözümlemesinin yapılmasına gerek olmasa da, tümce genelinde yapılması gereken bazı işlemler bulunmaktadır. Örneğin Türkmencede bazı kipler (gelecek zaman, zorunluluk ve planlama kipleri) şahıs eki almazlar, ancak Türkçede bu eylemleri yapan şahsın belirtilmesi zorunludur:

Türkmence	Türkçe
men geljek	geleceğ+im
sen geljek	gelecek+sin
o geljek	gelecek+Ø

Hedef dildeki tümcenin sözdizimsel çözümlemesinin olması durumunda özneye bakarak eyleme uygun şahıs bilgisini iliştirmek son derece basit bir işlemdir. Ancak doğrudan aktarım temel alındığı için sözdizimsel çözümleme yapılmamaktadır. Sistemin başarısının yükseltilebilmesi için, sözdizimsel çözümleme yapılmasa dahi, tümce bazında bazı basit işlemlerin gerçeklenmesi gereklidir. Örnek olarak yukarıda anlatılan, eksik olan şahıs bilgisinin çıkarılması sorununun kısmen çözülmesi için tümcenin başında bazı adılların varlığı sorgulanabilir.

Tümce genelinde yapılması gereken işlerden bir diğeri de Türkmencedeki ortaç öbeklerinde ortaya çıkmaktadır. Türkçede ortaçların kendisine gelen iyelik eki, Türkmencede ortacın nitelediği ismin sonuna gelmektedir:

Türkmence	Türkçe
berjek çöregi	vereceği ekmek
geljek yoluñyz	geleceğiniz yol

Ortaçlarla kurulan ad öbekleri, sözdizimsel çözümleme kadar kesin olmasa da yaklaşık olarak kestirilmeli ve iyelik etiketinin yeri değiştirilmelidir. Örnekleri verilen bu ve buna benzer durumların doğru olarak Türkçeye aktarılabilmesi için tümce genelinde sözcüklerin yapısal özellikleri üzerinde işlem gören kurallar tanımlanmıştır.

Türkçe biçimbilimsel üretici
Türkçe için geliştirilmiş geniş kapsamlı ve oturmuş bir biçimbiçimsel çözümleyici bulunmaktadır (Oflazer 1995). Biçimbilimsel çözümleyici SDD yapısında tasarlanmış olduğu için ters yönde çalıştırıldığında biçimbilimsel üretici olarak görev görmektedir.

Yüzeysel düzeyde çalışan tümce kuralları
Tümce genelinde çalışması gereken kuralların bazıları, sözcüklerin yapısal biçimleri yerine yüzeysel biçimlerine gereksinim duyarlar. Bu yüzden tümce geneli çalışan kuralların bir kısmı, Türkçe biçimbilimsel üretim aşamasından sonra çalıştırılmaktadır. Örneğin, Türkçede ayrı yazılan ama kendinden önceki sözcüğün son seslisine göre değişen ***de*** / ***da*** bağlacı ile ***mi*** / ***mı*** / ***mü*** / ***mu*** soru eklerinin doğru yüzeysel biçimlerinin oluşturulması ancak önceki sözcüklerin yüzeysel biçimleri üretildikten sonra olanaklıdır.

Başarımın ölçülmesi

Eğitim ve sınama verileri
İDM'lerin eğitilmesi için, yaklaşık 1 milyon sözcükten ve 50 bin tümceden oluşan bir derlem kullanılmıştır. Ayrıca Türkmence ile Türkçe arasındaki farklılıklardan dolayı derlem üzerinde uyumlaştırıcı bazı yapay değişiklikler yapılmıştır. Bunlardan birincisi, ismin aracılık durumu ile ilgilidir. Türkmencede, sözcüğe bitişik yazılan ***+ylA*** aracılık durum eki olmadığı için, Türkçe eğitim derleminde (***+Ins)*** durumunda olan tüm isim soylu sözcükler yalın hale (***+Nom***) getirilmiş ve bu sözcükten hemen sonra ***ile+Postp+PCNom*** eklenmiştir.

Derlem üzerinde yapılan ikinci değişiklik ise soru ekleri üzerindedir. Soru eki Türkmencede eyleme bitişik yazılmaktadır. Ancak Türkçede ayrı yazılan ve derlemde ***+mı, +mu, +mi, +mü*** şeklinde geçen farklı soru kökleri bulunmaktadır. Eğitim derleminde bu girdiler silinmiş, bunlardan önceki eylemin sonuna ***+Ques*** takısı eklenmiştir.

Sistemin başarısının sınanabilmesi için bir sınama kümesi oluşturulmuştur. Bu sınama kümesi, ağırlıklı olarak hikayelerden alınmış 255 adet Türkmence tümce ve bu tümcelerin 2 farklı kaynaktan sağlanmış Türkçe karşılıklarını (referans çevirilerini) içermektedir. Sınama kümesinin çevirisi sırasında oluşan biçimbilimsel ve sözcüksel belirsizliklerin oranları, Tablo 3'de gösterilmiştir.

Tablo 3. Belirsizlik dereceleri

Belirsizlik Türü	Oranı
Türkmencede Sözcük Başına Düşen Biçimbilimsel Çözümleme Adedi	1.55
Türkçe Karşılığı 1 Tane Olan Türkmence Sözcük Oranı	%44.7
Türkçe Karşılığı 2 Tane Olan Türkmence Sözcük Oranı	%34.0
Türkçe Karşılığı > 2 Olan Türkmence Sözcük Oranı	%21.3

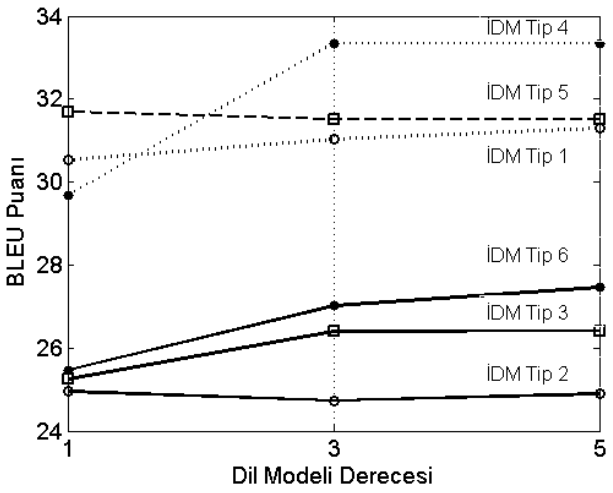
Başarım ölçütü

Önerilen çeviri modellerinin ve İDM türlerinin başarımlarını ölçmek için BLEU ölçütü kullanılmıştır (Papineni vd., 2002). Ölçütün temeli, sistem çıktısı aday tümcelerin, çevirmenler tarafından insan gücü ile çevrilmiş k adet referans çeviri ile olan benzerliğinin ölçülmesine dayanır. Benzerliğin ölçülmesi ise, sistem çıktısındaki sözcüklerin (1-gram) ve sözcük dizilerinin (2,3,4,…-gram), referans çevirilerdeki sözcük ve sözcük dizileri ile eşleştirilmesiyle yapılır. Uygulamada dörtten uzun sözcük dizilerinin eşleştirilmesinin gereksiz olduğu görülmüştür. Çevirinin doğası gereği, bir tümcenin aynı anlamı taşıyan birden fazla çevirisi olabilir. Sözcük ve anlatım biçimi seçimlerindeki bu serbestlik derecesi, değerlendirme aşamasında birden fazla referans çeviri kullanılarak çözülmeye çalışılmıştır.

Yüzeysel biçimlerin eşleşmesi temeline dayandığı için, bitişken diller söz konusu olduğunda BLEU yönteminin kullanışlılığı azalmaktadır. Değerlendirme aşamasında, bu olumsuzluğu düzeltmek amacıyla sadece köklerin eşleşmesi dikkate alınarak hesaplanan BLEU dereceleri de kullanılmıştır. BLEU sonuçları, sistem üzerinde yapılan değişiklikleri ve İDM'lerin başarımlarını ölçmek üzere kullanılmış, farklı sistemlerin sonuçları ile karşılaştırma amaçlı verilmemiştir (Callison-Burch vd., 2006).

Sonuçlar

Aktarım fonksiyonunun ürettiği tümcelerden en yüksek olasılığa sahip olan tümceyi bulmak için değişik İDM tipleri kullanılmış, üretilen çevirilerin BLEU ve BLEUr puanları hesaplanmıştır. İDM derecesinin etkisinin de görülebilmesi amacı ile n=1, 3 ve 5 seçilerek deneyler tekrarlanmıştır. Sonuçlar Şekil 3'de çizelge olarak verilmiştir.



Şekil 3. İDM başarımları çizelgesi

Çizelgeden aşağıdaki sonuçlara varılabilir:

1. İDM derecesi arttıkça, başarım artmaktadır.
2. $n \geq 3$ için İDM Tip 4 en yüksek başarımı sağlamaktadır.

Sonuçta, İDM Tip 4 ve n=3 seçilerek kullanılan BÇ sistemi, sınama veri kümesinden ürettiği aday çevirilerle, BLEU derecesi olarak 33, BLEUr derecesi olarak da 38 elde etmeyi başarmıştır.

Çeviri örnekleri

Aşağıda, sistemin ürettiği bazı çeviri örnekleri bulunmaktadır [2].

```
K: bu mesele bilen adamlar gyzyklanypdyrlar
   , emmä ol soraga her taraply jogap
   berip_bilmändirler .
```

[2] Örneklerde K: Kaynak, A:Aday (sistem çıktısı), R1 ve R2 ise sırasıyla 1. ve 2. referans çevirileri göstermektedir

```
A: bu meseleyle insanlar ilgilenmişler
   , ama o soruya her yönlü cevap    vere-
memişler .

R1:bu mesele ile insanlar ilgilenmişlerdir
   , ama bu soruya her yönü ile cevap
   verememişlerdir .

R2:bu meseleyle insanlar ilgilenmişlerdir
   , ama bu soruya her yönü ile cevap
   verememişlerdir .
```

Kaynak tümcede "_" işareti ile birleştirilen sözcükler, uygulama tarafından bir ÇSG olarak işaretlenmiş ve bu sayede Türkçeye doğru olarak çevrilebilmiştir. Aday çeviri ve referans karşılaştırıldığında anlamsal açıdan çok büyük bir farklılık olmamasına karşın, bu tümcenin BLEU puanının 25.82 çıkması, başarım puanlamasında sözcüklerin yüzeysel biçimlerinin eşleşmelerinin ne kadar önemli bir etkiye sahip olduğunu açıkça göstermektedir. Benzer şekilde, çeviri ile referans arasında hemen hemen hiçbir anlam farkı içermeyen aşağıdaki örnek için dahi BLEU puanı 56.69 olarak hesaplanmıştır:

```
K: bu dana adam birden gülmesini
   goÿup aglamaga başlapdyr.

A: bu bilgin adam birden gülmesini    bıra-
karak ağlamaya başlamış.

R1:bu bilgin insan birden gülmesini    ke-
sip ağlamaya başlamış

R2:bu bilgin insan birden gülmesini    bı-
rakıp ağlamaya başlamış.
```

Değerlendirmeler

Bu çalışmada, Türk dilleri arasında BÇ gerçeklemek üzere hem bilgi tabanlı ve kural temelli hem de istatistiksel bileşenlerden oluşan karma bir BÇ modeli önerilmiştir. Önerilen modelin başarımının ölçülmesi amacıyla Türkmenceden Türkçeye bir BÇ sistemi uygulaması gerçeklenmiştir. Geliştirilen sistem ile sadece Türkmenceden Türkçeye değil, diğer Türk Dilleri arasında BÇ gerçeklemek için de kullanılabilecek bir altyapı ortaya konulmuştur. Sistemin ürettiği çıktıların, başarılı BLEU dereceleri elde ettiği gösterilmiştir. Ancak bitişken diller için BLEU yönteminin eksiklikleri olduğundan, üretilen çıktılar üzerinde yapılan incelemelerde anlaşılabilirlik düzeyinin aslında daha fazla olduğu ortaya çıkmaktadır.

Sistemin diğer yönde yani Türkçeden Türkmenceye çeviri yönünde çalıştırılabilmesi için istatistiksel yöntemlerin eğitilebilmesi amacıyla Türkmence eğitim derleminin olması yeterlidir. Sistemin diğer tüm bileşenleri SDD olarak tasarlandığından, iki yönlü de çalışabilmektedir.

Geliştirdiğimiz bu alt yapı ile diğer Türk dilleri arasında çeviri yapılabilmesi için kaynak dile ilişkin biçimbilimsel çözümleyici ve hedef dile ilişkin biçimbilimsel üretici gereklidir. Bu araçlar, sadece geliştirilecek BÇ sistemine özgü bir gereksinim olmayıp, Türk dilleri üzerinde yapılacak neredeyse bütün doğal dil işleme çalışmaları için gereklidir. Aktarım için gerekli sözlüklerin hazırlanması, biçimbilimsel aktarım kurallarının çıkartılması da yapılması gereken diğer işlerdir. Son olarak, belirsizlik giderici bileşen için hedef dilde bir eğitim derlemi hazırlanarak İDM oluşturulmalıdır.

Sonuç olarak, Türk Dilleri arasında BÇ için doğrudan aktarım ilkesini temel alarak önerdiğimiz karma model ile geliştirilecek sistemlerin, başarılı çeviriler üretebileceği gösterilmiştir.

Kaynaklar

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

Altıntaş, K., ve Çiçekli, İ., (2001). A Morphological Analyser for Crimean Tatar, *Proceedings of the 10th Turkish Symposium on Artificial Intelligence and Neural Networks, TAINN*, North Cyprus.

Callison-Burch, C., Osborne, M., ve Koehn, P., (2006). Re-evaluating the Role of BLEU in Machine Translation Research, *Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, Trento, Italy.

Canals-Marote, R., Esteve-Guillén, A., Garrido-Alenda, A., Guardiola--Savall, M. I., Iturraspe-Bellver, A., Montserrat-Buendia, S., Pérez-Antón-Rojas, P., Ortiz-Pina, S., Pastor-Antón, H., ve Forcada, M. L., (2000). interNOSTRUM: a Spanish-Catalan Machine Translation System, *Machine Translation Review*, **11**, 21-25.

Hajič, J., (1987). RUSLAN - An MT System Between Closely Related Languages, *Third Conference of the European Chapter of the Association for Computational Linguistics (EACL'87)*, Copenhagen, Denmark.

Hajič, J., Homola, P., ve Kuboň, V., (2003). A simple multilingual machine translation system, *MT Summit IX*, New Orleans, USA.

Hajič, J., Hric, J., ve Kuboň, V., (2000). Machine Translation of Very Close Languages, *Proceedings of the Sixth Conference on Applied Natural Language Processing*

Hakkani-Tür, D. Z., Oflazer, K., ve Tür, G., (2002). Statistical Morphological Disambiguation for Agglutinative Languages, *Computers and the Humanities*, **36**, 381-410.

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.

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

Oflazer, K., Çetinoğlu, Ö., ve Say, B., (2004). Integrating Morphology with Multi-word Expression Processing in Turkish, *The ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, Barcelona, Spain.

Papineni, K., Roukos, S., Ward, T., ve Zhu, W. J. J., (2002). BLEU : A Method for Automatic Evaluation of Machine Translation, *Association of Computational Linguistics, ACL'02*, Philadelphia, PA, USA.

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

Tantuğ, A. C., Adalı, E., ve Oflazer, K., (2006b). Lexical Ambiguity Resolution for Turkish in Direct Transfer Machine Translation Models, *Lecture Notes in Computer Science, Volume 4263*, 230-238.