Türkçe ve Doğal Dil İşleme

Turkish Natural Language Processing

Kemal Oflazer
Carnegie Mellon Üniversitesi - Katar
Doha, Katar
ko@cs.cmu.edu

Özet

Bu makalede Türkçe'nin doğal dil işleme açısından ilginç olan özellikleri, ve karşılaşılan sorun ve bulunan çözümlerin kuş bakışı bir taraması yapılmıştır. Çoğu zorluklar dilin karmaşık sözcük yapısından ve bu yapının sözdizim ve istatistiksel modellemeyle olan ilişkisinden kaynaklanmaktadır. Bu taramanın sonrasında da Türkçe doğal dil işleme için geliştirilmiş olan önemli kaynakların bir özeti verilmiştir.

1 Giriş

Türkçe Altay dillerinin Türk dilleri ailesine giren bir dildir. Altay dillerinde Türk dillerinin dışında Moğol, Tunguz, Kore ve Japon dil aileleri de bulunur. Çağdaş Türkçe, Türkiye, Ortadoğu ve bazı Batı Avrupa ülkelerinde yaklaşık 60 milyon kişi tarafından anadili olarak konuşulmaktadır. Türk dilleri ailesinde bazıları ölü olan yaklaşık 40 dil vardır ve bu diller çok daha geniş bir coğrafyada yaklaşık 165-200 milyon kişi tarafında anadili olarak konuşulur (Bak. Şekil 1).[1] Tablo 1 de Türk dilleri ailesi içindeki dillerin önde gelenlerinin konuşanlarının oranlarını listelemektedir.[2]

Tablo 1: Türk dillerini konuşanların oranları

Dil	**%**
Türkçe	30.3
Azerice	11.7
Özbekçe	10.2
Kazakça	4.3
Uygurca	3.6
Tatarca	2.2
Türkmence	1.3
Kırgızca	1.0
Diğerleri	35.4

Türkçe ve Türk dilleri ailesinin diğer dillerinin doğal dil işleme açısından çok ilginç zorluklar içeren bir dizi özellikleri vardır. Türkçe, dilbilim ders kitaplarında özellikle eklemeli biçimbirim yapılı diller, ünlü uyumu veya tümce öğelerinin serbestçe yer değiştirilebilmesi konuların anlatıldığı zaman bu özelliklere sahip bir dil olarak örnek verilir.

Bu makalede Türkçe'nin doğal dil işleme açısından çok ilginç olan özelliklerine kuş bakışı olarak bir baktıktan sonra, Türkçe için geliştirilen doğal dil işleme teknikleri, sistemleri ve çeşitli kaynaklar hakkında özet bilgiler vereceğiz.[3]

[1] Kaynak: Vikipedi
[2] Kaynak: Vikipedi
[3] Dilbilim açısından Türkçe ile ilgili bilgi almak isteyenlere Kerslake ve Göksel'in kitabını öneririz [13].

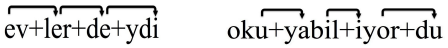
Şekil 1: Türk dillerinin coğrafyası

2 Türkçe'nin Biçimbilimsel Yapısı

Biçimbilim açısından Türkçe bitişken bir dildir; biçimbirimler bir kök sözcüğe "tespih taneleri" gibi eklenirler. Türkçe'de önek yoktur; ayrıca üretken olarak, örneğin, Almanca'daki gibi bir dizi isim kökü birbirine ekleyerek beraber yazılan birleşik isimler de bulunmaz. Birleşik isimler genellikle öğelerinin anlamlarının toplamından çok daha farklı anlamlar için kullanılan "sözcükleşmiş" birleşik isimlerdir (örn.: *acemborusu*, bir çiçek adıdır.)

Türkçe'de sözcükler yaklaşık 30 bin kadar kök sözcüğe çok üretken bir şekilde bir dizi ek ekleyerek oluşturulur.[4] İsimler örneğin Almanca veya Fransızca'da olduğu gibi sınıflara ayrılmaz. Sözcük dağarcığı tarihsel, coğrafi ve ekonomik nedenler yüzünden zaman içinde Arapça, Farsça, Yunanca, Ermenice, Fransızca, İtalyanca, Almanca ve son 50-60 sene içinde de İngilizce'den etkilenmiştir.

[4] Özel isimleri saymamaktayız.



Şekil 2: Ünlü uyumunun silsile şeklinde çalışmasının iki örneği

Bir tümce içinde kullanıldığında sözcükler bir dizi çekim ve yapım eki alır. Örneğin bir sözcük İngilizce'de ifade edildiğinde bir tümceye karşılık gelebilir:

yap+abil+ecek+se+k →
if we will be able to do (it)

Hemen hemen tüm biçimbirimlerin kullanılan ünlüler ve biçimbirim sınırlarındaki ünsüzler yönünden farklı şekilleri vardır; örneğin:

paket+ten araba+dan

Burada solda, ekin ilk ünsüzü ve ünlüsü, bitiştiği gövdenin son ünsüz ve ünlüsü ile uyum içinde olmak için *t* ve *e* olarak seçilirler. Sağdaki gövde, ünlü ile bittiği için ekin ilk ünsüzü *d* olarak kalır ancak ünlü uyum için *a* olmak durumundadır. Ünlü uyumu dediğimiz bu süreç Şekil 2'de görüldüğü gibi soldan sağa doğru silsile olarak gider. Türkçe sözcüklerin iki düzeyli biçimbilim çerçevesinde biçimbirimlerine ayrıştırılmasının detayları için Oflazer'e [16] başvurabilirsiniz.

Türkçe sözcüklerde yapım eklerinin bulunmasına sıklıkla rastlanır. Böyle sözcükler bazen çok karmaşık yapıya sahip olabilirler. Ebru Arısoy doktora tezinde [2] derlediği çok büyük bir derlemde rastladığı dokuz biçimbirimli *ruhsatlandırılamamasındaki* sözcüğünü örnek olarak vermektedir. Bu sözcüğün iç yapısında beş adet yapım eki vardır: Şekil 3'te görüldüğü gibi sözcük kökten isim olarak başlayıp 5 türetme sonrasında bir sözcük haline dönüşmektedir

Ancak istatistiksel olarak durum bu sözcükte olduğu gibi kötü değildir. Genelde bir derlemdeki sözcüklerde ortalama yaklaşık üç biçimbirim gözlenir. Ancak bu biraz yanıltıcıdır çünkü yüksek sıklıkta görülen sözcükler genelde tek bir biçimbirimden

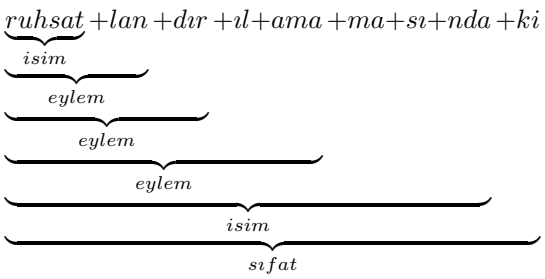


Figure 3: Karmaşık bir Türkçe sözcükteki türetmeler

oluşur. Ayrıca sözcüklerin biçimbilim yapısı açısından ortalama iki farklı yorumu vardır – bunlar kök sözcüğün sınıfının farklı olması (örneğin isim *ek* veya eylem *ek*), sözcüğün birimbirimlere farklı şekillere bölünmesi (*oku+ma* veya *ok+um+a*), aynı yazılan biçimbirimleri farklı anlamlara gelmesi gibi nedenlerden oluşur (emir kipi *oku+ma* veya mastar *oku+ma*).

Tablo 2 büyük bir Türkçe derlemdeki en sık yirmi sözcüğü, yanlarında biçimbirim sayısı ve farklı yorum sayısı ile birlikte göstermektedir. Bu rakamlarda kabaca şu sonucu çıkarabiliriz: (i) yüksek sıklıktaki sözcüklerin çoğunda bir biçimbirim olduğuna göre, ortalamanın üç olması için düşük sıklıktaki sözcüklerin üçten daha fazla birimbirimi olacaktır; (ii) ayrıca yine yüksek sıklıktaki sözcüklerin çoğunun çok sayıda yorumu olduğuna göre ortalamanın iki yorum olması için düşük sıklıktaki sözcüklerin genelde ortalama ikiden az yorumu olması beklenebilir.

Türkçe sözcüklerin biçimbilimsel yapılarının bir diğer önemli özelliği de daha önce de vurguladığı gibi, yapım eklerinin çok sık kullanılmasıdır. Tablo 3'de tek bir isim (*masa*) ve eylem (*oku*) kökünden 0, 1, 2, 3 yapım eki kullanılarak elde edilebilecek farklı sözcüklerin sayısı görülmektedir.[5,6] Tabii ki burda sayılan sözcüklerin çoğu hiç kullanılmayabilir, ama bu sayılar en azından dilin biçimbilim yapısının üretim gücünü gösterir. Tek bir eylem kökünden nerdeyse 1.5 milyon değişik sözcük üretilebilmesi hayret edilecek bir özelliktir (Bunun üzerine oldukça eğlenceli ve ilginç bir çalışma olarak Wickwire'ın tezini önerebiliriz [26].)

Tablo 3: 0, 1, 2, 3 yapım eki ile tek bir Türkçe isim veya eylem kökünden üretilebilecek sözcük sayısı

Kök	Yapım Eki	Sözcük	Toplam
masa	0	112	112
	1	4,663	4,775
	2	49,640	54,415
	3	493,975	548,390
oku	0	702	702
	1	11,366	12,068
	2	112,877	124,945
	3	1,336,266	1,461,211

Bu üretkenlik gerçek kaynaklardan toplanan derlemlere de yansır. Sak ve diğerleri [24], yaklaşık 500 milyon sözcüklük bir haber derleminden topladıkları istatistiklerde şu gözlemlere varmışlardır: Bu derlemde toplam 4.1 milyon farklı sözcük vardır ve bunların en sık geçen 50 bini derlemin %89'unu, en sık geçen 300 bini ise %97'sini kapsamaktadır. 3.5 milyon sözcük 10 defadan az geçmektedir, 2 milyon sözcük ise sadece bir kere geçmektedir. Fakat en can alıcı gözlem ise şudur: Derleme 490 milyon sözcükten sonra 1 milyon sözcük daha eklenince daha önceden hiç karşılaşılmamış 5,539 yeni sözcük gözlenmiştir. Bu gözlemi örneğin İngilizce bir derlemde yapmak olası değildir. Yine aynı çalışmada bu derlemdeki sözcükler kök ve kök sonrası ek dizisi olarak ayrılır ve her bir gruptaki farklı kökler ve ek diziler sayılırsa belli bir noktadan sonra (yaklaşık 360 milyon sözcük), karşılaşılan farklı ek dizilerinin sayısı farklı köklerin sayısını geçmektedir (Bakınız Şekil 4.) Bu pratikte sonsuz addedilecek sözcük dağarcığı hemen hemen her türlü doğal dil işleme uygulamasında ilginç sorunlar çıkarmaktadır.

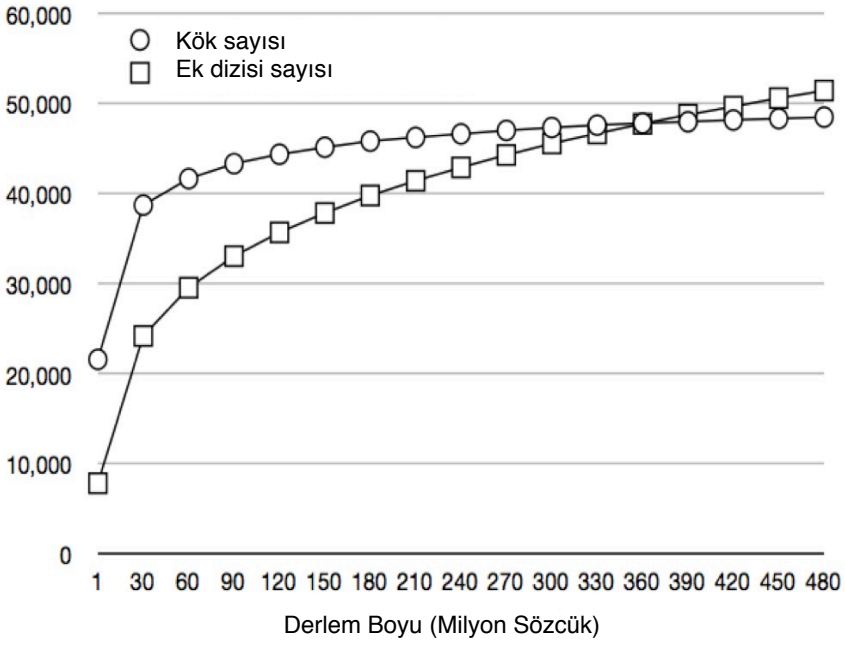
Yazım Düzeltmesi: Yazım düzeltmesi için diğer diller için geliştirilen ve sonlu bir sözcük dağarcığı

[5] 0 yapım eki sadece çekim ekleri ile oluşturulabilen sözcüklerin sayısına karşılık gelir.

[6] Bu sayılar Xerox'un *xfst* yazılımı ile Türkçe biçimbirim çözümleyicisinden çıkarılmıştır.

Tablo 2: En sık yirmi sözcük, biçimbirim sayıları ve farklı yorum sayıları

	Sözcük	Biçimbirim	Yorum		Sözcük	Biçimbirim	Yorum
1	bir	1	4	11	kadar	1	2
2	bu	1	2	12	ama	1	3
3	da	1	1	13	gibi	1	1
4	için	1	4	14	olan	2	1
5	de	1	2	15	var	1	2
6	çok	1	1	16	ne	1	2
7	ile	1	2	17	sonra	1	2
8	en	1	2	18	ise	1	2
9	daha	1	1	19	o	1	2
10	olarak	2	1	20	ilk	1	1



Şekil 4: Farklı köklerin ve ek dizilerinin derlem boyu ile artması

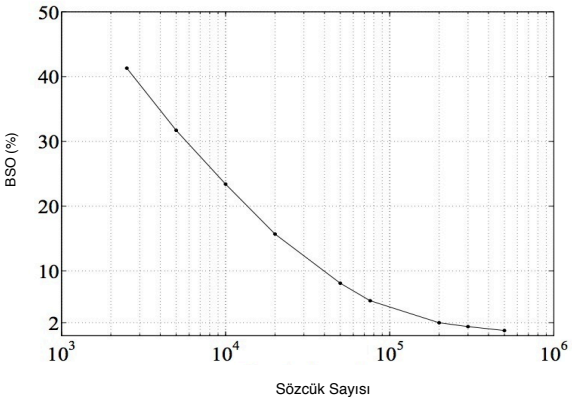
kabulune dayanan teknikler Türkçe için uygun değildir. Önceki çalışmalarımızda [17], Türkçe gibi diller için sözcük dağarcığını sonlu durumlu bir dönüştürücü ile göstermeye dayanan ve sonlu durumlu makina çizge yapıları üzerinde çok etkin bir şekilde hataya dayanıklı yaklaşık arama yapan algoritmalar geliştirdik.

İşaret Kümesi Tasarımı: Türkçe sözcüklerin biçimbilimsel çözümlemesi sonucunda çıkan bilgileri İngilizce veya Almanca gibi dillerde olduğu gibi sonlu sayıda işaret ile göstermek olanaklı değildir. Şekil 4'teki gözlem zaten bunun için ipuçları vermektedir (yani eklerde kodlanan bilgi kuramsal ve pratik olarak sonlu bir sınır içinde değildir.) Her ne kadar az sayıda sözcük sınıfı olsa da yapım ve çekim eklerinin sayısının önceden belirli bir sayıda olmaması, işaret sayısının sonlu olmasını önlemektedir. Türkçe sözcükler için gözlemlenen işaret sayıları hakkında istatistikler için Hakkani-Tür ve diğerlerine [12] bakmanızı öneririz.

İstatistiksel Dil Modelleme: Büyük sözcük dağarcığı istatistiksel dil modellemede hemen hemen her zaman veri yetersizliği probleminin yaşanmasına neden olur. Ebru Arısoy'un doktora tezinden [2] alınan Şekil 5, bir konuşma tanıyıcı sisteminde kullanılan dil modeli için farklı sözcük sayısına göre bilinmeyen sözcüklerin oranı hakkında bir bilgi vermektedir Yine aynı tezden alınan Tablo 4 also yaklaşık 60 bin sözcüklük bir sözcük dağarcığı ile test kümesinde rastlanan bilinmeyen sözcüklerin yüzdesini göstermektedir. Görüleceği gibi Türkçe ve çekimli bir dil olan Çekçe'de %8 gibi bir bilinmeyen sözcük oranı vardır. Türkçe gibi bitişken diller olan Fince ve Estonyaca'da ise çok daha yüksek oranlar gözlenmiştir.

Arısoy ayrıca dil modelleme için biçimbirimleri kullanmış ve de yaklaşık 76 bin kök ve biçimbirim kullanarak test kümesi için çok çok düşük bir bilinmeyen sözcük oranı gözlemiştir.

Sözdizim modellemesi: Aşağıda göreceğimiz gibi, yapım ekleri sözdizim modellemesi açısından çok ilginç işlevlere sahiptirler. Bu özellikler hem



Şekil 5: Dil modellemede sözcük dağarcığı ile bilinmeyen sözcüklerin oranı (BSO) ilişkisi

Tablo 4: Birkaç dil için bilinmeyen sözcüklerin oranı (BSO)

Dil	**Vocab.**	**BSO**
İngilizce	60K	1%
Türkçe	60K	8%
Fince	69K	15%
Estonyaca	60K	10%
Çekçe	60K	8%

öbek tabalı modellemeler hem de bağımlılık tabanlı modellemelerde geçerlidirler. Biçimbirim – sözdizim etkileşimi için Çetinoğlu ve Oflazer [7] ve Eryiğit ve diğerlerine [11] bakmanızı öneririz.

İstatistiksel Çeviri: İstatistiksel dil modellemede olduğu gibi istatistiksel çeviride de büyük sözcük dağarcığı veri yetersizliği problemini öne çıkarır. Bu problemi aşmak için yine biçimbilimsel yapıya dayalı çeviri yaklaşımları oldukça iyi sonuçlar elde etmişlerdir.

3 Tümce Öğe Sırası ve Biçibirim–Sözdizim Arabirimi

Türkçe tümcelerde doğal öğe sırası *Özne - Nesne - Yüklem* şeklindedir – zaman, yer, vb. belirten diğer belirteç öğeler hemen hemen herhangi bir yere gidebilirler. Ancak *Özne - Nesne - Yüklem*'in diğer 5 sırası da gerekli durumlarda özellikle gerekli çevrimsel şartlarda da kullanılabilirler.[7] Öğe sırasının bu şekilde serbest olması diğer aynı özelliğe sahip dillerde olduğu gibi tümcedeki isim öbeklerinin baş sözcüklerinin işleve göre durum ekleri almasıyla sağlanır.

Aşağıdaki örnekler bu değişik temel öğe sıralarını ve her birisi için öngörülen çevrimsel kabulleri veya beklentileri göstermektedir. Her tümcede ana eylem Ekin'in Ayşe'yi görmesidir – sıra değişiklikleri konuşma sırasında çevrimi, kabul edilen önbilgileri ve beklentileri kodlamaktadır.

- Ekin Ayşe'yi gördü.
- Ayşe'yi Ekin gördü. (*gören Ekin'di başka birisi değil!*)
- Gördü Ekin Ayşe'yi. (*ama görmemesi gerekiyordu.*)
- Gördü Ayşe'yi Ekin. (*zaten görmesini bekliyordum!*)
- Ekin gördü Ayşe'yi. (*başkası da görebilirdi*)
- Ayşe'yi gördü Ekin. (*başkasını da görebilirdi!*)

Bu değişik öğe sıralarını geleneksel Çevrimden Bağımsız Gramer formalizmaları ile modelleme her ne kadar olanaklı ise de model beklendiği kadar temiz veya basit değildir. Çetinoğlu'nun doktora tezinde [6] geliştirdiği büyük boyutlu Sözcüksel İşlevsel Gramer temelli gramer bu sıra farklılıklarını oldukça prensipli bir şekilde nodellemiş olsa da sıra farklılıklarının getirdiği ek bilgileri kodlamak için mekanizmaların olmaması bunları kodlanmasına olanak vermemiştir.

3.1 Biçimbirim – Sözdizim Arabirimi

Sözcük yapılarının ve özellikle de yapım eklerinin sözdizim modellemede çok ilginç ilişkileri vardır. Bunun detaylarına girmeden önce bunu açıklamamızda yardımcı olacak bir soyutlamayı açıklamakta fayda vardır.

Türkçe'de bir sözcüğün biçimbirim yapısını en genel hali ile bir kök sözcüğe eklenen ve biçimbirimlerin içerdiği bilgiyi gösteren işaret dizileri ile kodlarız. Bu işaretlerden bir tanesi olan `^DB` yapım eklerinin sınırlarını gösterir. Sözcüğün başından ilk yapım ekine, son yapım ekinden sözcüğün sonuna, ve de iki yapım eki arasındaki çekim eklerinde oluşan her bir gruba *çekim grubu* (İngilizce yayınlarımızda kullandığımız adı ile *inflectional group*(IG)) adı vermekteyiz. Dolayısı ile çözümlerde her biri şu şekilde gösterilir:

$$\texttt{kök+IG}_1 + \texttt{\^{}DB+IG}_2 \cdots + \texttt{\^{}DB+IG}_n.$$

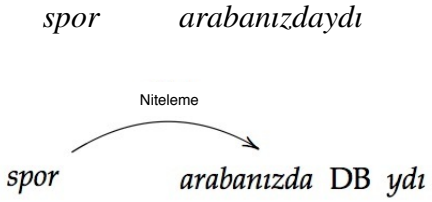
Burada her $\texttt{IG}_i$ kökün ve diğer yapım eklerinin sözcük sınıfları dahil olmak üzere tüm çekim bilgilerinden oluşur. Bir sözcüğün her biri bu şekilde gösterilen birden çok biçimbilimsel gösterimi olabilir. Bunların herbirisi sözcüğün biçimbirim yapısının farklı gösterimine karşılık gelir. Örneğin `uzaklaştırılacak` sözcüğünün gösterimi şu şekildedir:[8]

```
uzak+Adj
      ^DB+Verb+Become
      ^DB+Verb+Caus
      ^DB+Verb+Pass+Pos
      ^DB+Adj+FutPart+Pnon
```

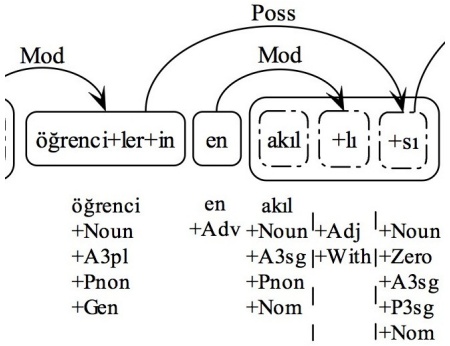
Bu gösterimdeki 5 çekim grubu şu şekildedir:
1. `+Adj`

[7] Sıklıkla öne çıkarılan bir kısıt tümcedeki yalın durumdaki belirtisiz nesnenin her zaman yüklemin hemen öncesinde olması gerektiğidir. Ancak bu kısıtın da geçerli olmadığı örnekler de gözlenmiştir (örneğim, *Yapayım sana bir yemek.* (Sarah Kennelly, özel konuşma).

[8] Kullanılan sembolleri Türkçe karşılıkları şu şekildedir: `+Adj`: Sıfat, `Verb`: Eylem, `+Become`: Dönüşüm yapım eki, `+Caus`: Ettirgen, `+Pass`: Edilgen, `+Pos`: Olumlu, `+FutPart`: Gelecek zaman ortacı, `+Pnon`: İyelik eki yok.



Şekil 6: Çekim grupları arasındaki ilişkiler



Şekil 8: Şekil 7'deki bir sözcüğün birden fazla ilişkisi

2. `+Verb+Become`
3. `+Verb+Caus`
4. `+Verb+Pass+Pos`
5. `+Adj+FutPart+Pnon`

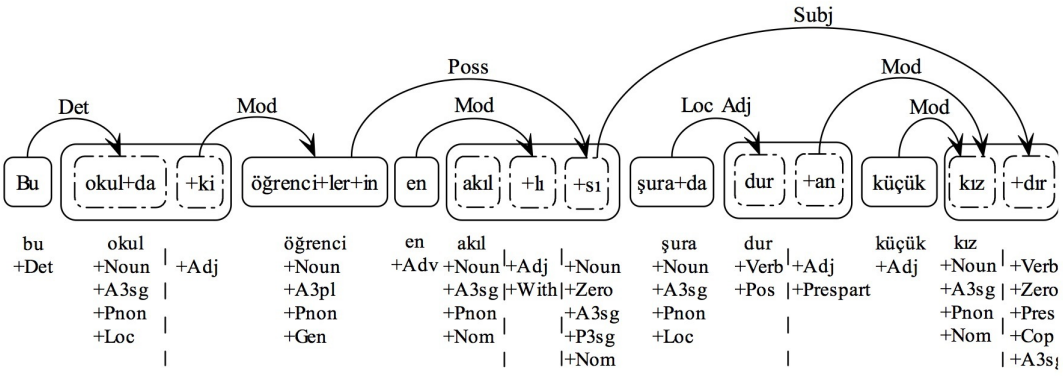
Birinci çekim grubu sadece kökün sıfat olduğunu belirtir. İkinci çekim grubu önceki sıfat kökünde anlamı o sıfata dönüşmek olan bir eylem türetir (*uzaklaş*). Üçüncü çekim grubu önceki eylemden ettirgengen bir eylem türetildiğini gösterir (*uzaklaştır*). Dördüncü çekim grubu ise bir öncekinden edilgen bir eylem türetidiğini gösterir (*uzaklaştırıl*). En son olarak da bir önceki edilgen eylemden bir gelecek zaman ortacı türetilir ki bu da tümcede bir başka isim öbeğinin niteleyicisi olarak kullanılacaktır.

Çekim gruplardan bahsetmemizin en önemli nedeni tümce içindeki sözdizimsel ilişkileri sözcükler arasında değil de sözcüklerin parçaları olan çekim grupları arasında olmasıdır. Ayrıca bir sözcüğün tümce içindeki işlevi sadece son çekim grubunun çekim özellikleri tarafından belirlenir. Bunun için Şekil 6'daki çok basit örneği verebiliriz. *Spor arabanızdaydı* tümcesinde ikinci sözcüğün ikinci çekim grubu türetilmiş bir eylemdir ve bu tümcenin yüklemi işlevini görmektedir. Öncesinde ise ilk sözcük ve ikinci sözcüğün ilk çekim grubundan oluşan bir isim tamlaması vardır – yani *spor* sözcüğü *araba* ile ilişkilidir ve onu niteler; tümcenin yüklemi olan kısım ile bir ilişkisi yoktur. En genel durumda bir sözcüğün çekim gruppları farklı sözcüklerin çekim grupları ile faklı ilişkiler içinde olabilirler. Bunun için Şekil 7'de görülen ve Ağaç Yapılı Türkçe Derlem'deki tümceleri nasıl kodlandığını da gösteren örneğe bakılmasını öneririz.[9] Bu şekilde düz çizgili oval dörtgenler sözcükleri ve kırık çizgili oval dörtgenler ise çekim gruplarını göstermektedir. Önce de söylediğimiz gibi ilişki oku her sözcükte son çekim grubundan çıkmakta ve (genelde sağda bulunan) başka bir sözcüğün çekim gruplarından bir tanesine gitmektedir. Her çekim grubunun biçimbilimsel özellikleri dikey olarak altında listelenmiştir. Örneğin tümcenin ortasındaki üç sözcüğe odaklanırsak (Şekil 8) şunları görebiliriz:

- *akıllısı* sözcüğü üç çekim grubundan oluşmaktadır: *akıl* isminden *+lı* eki ile sıfat türetilmiş, hemen akabinde de bunda tekrar isim üretilmiştir.
- *öğrencilerin* sözcüğü ve de *akıllısı* sözcüğünün son çekim grupları belirtili isim tamlaması kurmak için gerekli biçimbilimsel özelliklere sahiptirler ve şekilde *Poss* ile belirtilen ok bu ilişkiyi gösterir.
- Aradaki *en* sözcüğü ise *akıllısı* ikinci çekim grubu olan sıfat ile ilişkilidir – *en* belirteci sadece bir sıfatla ilişkiye girebilir.

Çekim grubu kavramını daha önceki çalışmalarımızda gösterimi ve modellemeyi kolaylaştıran bir soyutlama olarak kullandık: Hakkani-Tür diğerleri

[9] Burada sadece yüzeysel bağımlılık ilişkilerini göstermekteyiz ve ilişki okları bağımlı birimden baş birime gitmektedir.



Şekil 7: Bir tümcedeki çekim grupları arasındaki ilişkiler

[12] istatistiksel modellemede çekim gruplarını kullandı. Çetinoğlu doktora tezinde [6] çekim gruplarını Türkçe için sözcüksel işlevsel gramer geliştirirken kullandı. Eryiğit ve diğerleri [11] Türkçe için bağımlılık çözümlemesi yapmak için yine çekim gruplarını kullandı. Ağaç Yapılı Türkçe Derlem de [21] çekim grupları arasındaki ilişkileri kodladı.

4 İstatistiksel Çeviri

Bu noktada Türkçe'nin biçimbilimsel yapısının istatistiksel çeviri sistemleri için de sorun olacağı açıktır. Bunu daha da vurgulamak için de belki biraz zorlama olarak görülebilecek, ama çok da anlamsız olmayan şu örneği, İngilizce bir tümcenin bir kısmının nasıl Türkçe'ye dönüştürülebileceğine sürecine bir örnek olarak verebiliriz. Şekil 9 bu varsayımsal ve ideal çeviri sürecini göstermektedir. İngizce sözcükler önce doğru yerlere kaydırılır, sonra her biri gerekli Türkçe kök ve biçimbirimlere aktarılır ve sonra bunlar birleştirilip karşılık Türkçe sözcük oluşturulur.

Burdan hemen görebiliriz ki istatistiksel çeviri sistemlerinin eşleştirme öğrenme safhası için sözcük bazında eşleştirme yapmak çok sorunlu olacaktır. Türkçe tarafında da tek bir biçimbirim bile yanlış aktarılsa veya yanlış yere konsa tüm Türkçe sözcük yanlış olacaktır! Bu durumda ilk akla gelecek yaklaşım değişikliği, Türkçe sözcükleri biçimbirimlerine bölerek ve de biçimbirimlerine sanki birer sözcükmüş muamelesi yaparak İngilizce tarafıyla eşleştirmektir. Bu durumda tümceler eşleştirmeye şu şekilde girer.

E: *I would not be able to do ...*
T: *... yap +ama +yacak +tı +m*

Bu yaklaşım Durgar-El Kahlout'un doktora tezinde [8], ve öncesi ve sonrasındaki çeşitli yayınlarda [18, 8, 9] Moses çeviri sistemi [14] kullanılarak denendi. Her ne kadar sözcük tabanlı bir sistemle karşılaştırıldığında oldukça iyi ilerlemeler kaydedilmiş olsa da başka bazı önemli problemler de gözlendi:

- Türkçe sözcükler biçimbirimlere ayrılınca ortalama "tümce boyu" nerdeyse 3 misline çıktı ve bu eşleştirme için ciddi sorunlar çıkardı.
- Görevi sadece sözcükleri doğru aktarıp doğru yerlerine yerleştirmek olan çözücü birimi ise bu gösterimle hem sözcüklerin doğru sırasını hem de sözcükler içindeki biçimbirimlerin doğru sırada çıkarılmasını sağlamak zorunda kaldı. Bu nedenle ciddi oranda sözcükte biçimbirim sırası yanlış olarak aktarıldı.

Farklı bir yaklaşım ise İngilizce tümcelerdeki belli sözdizimsel yapıları tanıyarak bunları Türkçe

Şekil 9: İngilizce nasıl Türkçe'ye dönüşür

sözcüklere benzetmeye dayalı oldu. Yeniterzi ve Oflazer [27] *sözdizim - biçimbirim* aktarması olarak adladırılan bu yaklaşım ile İngilizce tümcelerdeki çeşitli yapıların önce dönüştürülerek bunların Türkçe'deki sözcüklere benzemeleri sağlandı. Bu şekilde İngilizce tarafında çoğu işlev sözcüğü sanki biçimbirimmiş gibi başka sözcüklere iliştirildi. Mesela İngilizce tümcede şu şekilde bir öbek varsa

... in their economic relations ...

bir sözdizimsel çözümleyici *in* ilgecinin and iyelik adılı *their* sözcüklerinin *relations* sözcüğüne ilişkili olduğunu çıkarıp, bu öbeği söyle bir gösterime dönüştürürdü:

... economic relation+s+their+in ...

Türkçe eş tümcede de biçimbiçimlere ayrılınca elimize şu geçti

... ekonomik ilişki+ler+i+nde ...

Sonrasında da tümceler sadece *kök* sözcükler bazında eşleştirildi ve bu eşleşmede eşleşen kök sözcüklerin biçimbirim dizilerinin de eşleştiği kabul edildi. Bu dönüştürmeler sonucunda İngilizce tümcelerin boyu %30 azaldı ve eşletirme süreci çok daha sağlıklı oldu. Bu yaklaşımla da oldukça iyi sonuçlar elde edildi ve en önemlisi üretilen Türkçe sözcüklerin biçimbirimlerinin ve de bunların sıralarının yanlış olarak çıkmalarının önüne geçildi.

5 Türkçe Doğal Dil İşleme için Geliştirilen Kaynaklar

Geçtiğimiz yirmi yıl içinde Türkçe doğal dil işlemede kullanılabilecek bir dizi kaynak geliştirilmiştir. Bu bölümde bunların en önemlilerinin üzerinden kısaca geçip nereden edinilebileceğine dair bazı yönlendirmeler yapacağız.

1. *Biçimbilimsel Çözümleme*: Oflazer[16] çalışmasında Türkçe için iki düzeyli biçimbilim formalizması temelinde bir çözümleyicinin detayları görülebilir. Bu çözümleyici Xerox sonlu durumlu makinalar yazılımı ile geliştirilmiştir. Bu çözümleyici çok daha genel bir çözümleyici olarak aynı anda hem biçimbilimsel yapıyı hem de sesbirim, hece sınırı ve vurguyu da üreten bir sistem olarak da gerçekleştirilmiştir [19].

2. *Biçimbilimsel Tekleştirme*: Oflazer ve Kuruöz [20], Oflazer ve Tür [22], Hakkani-Tür ve

diğerleri[12] gibi eski çalışmalara ek olarak son zamanlarda Sak ve diğerleri [23] ve Yuret ve Türe [28] tarafından daha yeni yaklaşımlar kullanılarak pratikte oldukça iyi çalışan tekleştiriciler geliştirilmiştir.

3. *İstatistiksel Bağımlılık Çözümleyicisi* : Türkçe için, Ağaç Yapılı Derlem [21] ile eğitilmiş bir dizi bağımlılık çözümleyisi geliştirilmiştir. Eryiğit ve Oflazer [10] direk olarak çekim grupların arasındaki ilişkilerin istatistiklerine dayanan bir çözümleyiciyi tanıtır. Eryiğit ve diğerleri [11], ise MaltParser yaklaşımını [15] kullanan deterministik bir bağımlılık çözümleyiciyi anlatır.[10]

4. *Sözcüksel-İşlevsel Gramer Temelli Çözümleyici*: Geniş kapsamlı ve derin cözümleme yapabilen bir sistem Özlem Çetinoğlu tarafından doktora tezinde [6], ParGram (Parallel Grammars) Projesinin [5] içinde geliştirilmiştir.[11] Bu çalışmanın amacı dilbilimsel bir dizi özellik için belli bazı ilkelere dayanan ve diğer dillerdeki tümcelerin işlevsel çözümlerine yakın koşutlukta derin çözümler çıkaran bir çözümleyici elde etmekti.

5. *Ağaç Yapılı Derlem*: Türkçe için 5,635 tümcelik ve çekim gruplarına dayalı bir gösterim kullanan bir ağaç yapılı derlem geliştirilmiş ve araştırmacıların kullanımına açılmıştır [21].[12] Bu derlem başka bir dizi çalışmanın ötesinde, yakın geçmişte CONLL Çok Dilli Bağımlılık Çözümlemesi yarışmalarında [4], kullanılmıştır.

6. *Türkçe WordNet*: Balkanet projesi [25] çerçevesinde Türkçe için yaklaşık 15 bin eşanlamlılar kümesinden oluşan bir kavramsal sözlük geliştirilmiştir[3] ve çok sayıda araştırmacı tarafından araştırmalarda kullanılmıştır.

7. *Çeşitli diğer kaynaklar*: Bunlara ek olarak başka bir dizi kaynak da geliştirilmiştir veya geliştirilmektedir. Bunların arasında en önemlisi olarak Türkçe Ulusal Derlemi'ni gösterebiliriz [1] (Ayrıca bakınız `http://www.tnc.org.tr/`. Deniz Yuret ise `http://www.denizyuret.com/2006/11/turkish-resources.html` sitesinde Türkçe için bulunan kaynakların güncel bir listesini vermektedir.

6 Sonuçlar

Her ne kadar geniş bir coğrafyada 60 milyon kişi tarafından anadili olarak konuşulan bir dil olsa da Türkçe üzerindeki doğal dil işleme çalışmaları ancak son 15-20 yıl içinde hız kazanmıştır. Türkçe bir dizi özelliği nedeniyle dil işleme için çok ilginç bazı problemlere yol açmış olsa da bunlar için elde edilen çözümlerin yeterli şekilde soyutlandıklarında çok daha geniş bir dil kümesine de uyarlanabilir olduğu gözlenebilmiştir.

Her ne kadar zaman içerisinde Türkçe için biri dizi kaynak geliştirilmiş olsa da hala bazı engeller vardır: Örneğin istatistiksel çevirinin ana ham maddesi olan bir tarafı Türkçe olan koşut derlemler için doğal bir kaynak yoktur (mesela 20 dilde koşut olarak yazılan AB parlamentosu tutanakları gibi). Yine de bir sürü sıkıntıya rağmen son 10 yılda gerek Türkiye'de gerekse de dışarda, bu konu üzerinde çalışan araştırmacıları ve araştırma grupların yavaş da olsa artıyor olması ümit vericidir.

Kaynaklar

[1] Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Ümit Mersinli, Demirhan, U.U., Yılmazer, H., Atasoy, G., Öz, S., İpek Yıldız, Özlem Kurtoğlu: Construction of the Turkish National Corpus (TNC). In: N. Calzolari, K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds.) Proceedings of the Eight International Conference on Language Resources and Evalu-

[10] Bu çözümleyici `http://web.itu.edu.tr/gulsenc/TurkishDepModel.html` sitesinden indirilebilir.

[11] Ayrıca bakınız: `http://pargram.b.uib.no/`

[12] `www.ii.metu.edu.tr/corpus/treebank.html`. sitesinden indirilebilir.

ation (LREC'12). European Language Resources Association (ELRA), İstanbul, Türkiye (2012)

[2] Arısoy, E.: Statistical and discriminative language modeling for Turkish large vocabulary continuous speech recognition. Doktora Tezi, Boğaziçi Üniversitesi (2009)

[3] Bilgin, O., Çetinoğlu, O., Oflazer, K.: Building a Wordnet for Turkish. Romanian Journal of Information Science and Technology **7**(1-2), 163–172 (2004)

[4] Buchholz, S., Marsi, E.: CoNLL-X shared task on multilingual dependency parsing. Proceedings of CoNLL, Sayfa 149–164 (2006)

[5] Butt, M., Dyvik, H., King, T.H., Masuichi, H., Rohrer, C.: The parallel grammar project. Proceedings of the COLING-2002 Workshop on Grammar Engineering and Evaluation, Sayfa 1–7 (2002)

[6] Çetinoğlu, O.: A large scale LFG grammar for Turkish. Doktora Tezi, Sabancı Üniversitesi (2009)

[7] Çetinoğlu, O., Oflazer, K.: Integrating derivational morphology into syntax. In: N. Nicolov, G. Angelova, R. Mitkov (eds.) Recent Advances in Natural Language Processing. John Benjamins (2009)

[8] Durgar-El-Kahlout, I.: A prototype English-Turkish statistical machine translation system. Doktora Tezi, Sabancı Üniversitesi (2009)

[9] Durgar-El-Kahlout, I., Oflazer, K.: Exploiting morphology and local word reordering in English to Turkish phrase-based statistical machine translation. IEEE Transactions on Audio, Speech, and Language Processing **18**(6), 1313–1322 (2010)

[10] Eryiğit, G., Oflazer, K.: Statistical dependency parsing of Turkish. Proceedings of the 11th EACL, Sayfa 89–96. Trento, İtalya (2006)

[11] Eryiğit, G., Nivre, J., Oflazer, K.: Dependency parsing of Turkish. Computational Linguistics **34**(3), 357–389 (2008)

[12] Hakkani-Tür, D., Oflazer, K., Tür, G.: Statistical morphological disambiguation for agglutinative languages. Computers and the Humanities **36**(4) (2002)

[13] Kerslake, C., Göksel, A.: Turkish: A Comprehensive Grammar. Comprehensive Grammars. Routledge (Taylor and Francis), New York, ABD (2005)

[14] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Sayfa 177–180. Association for Computational Linguistics, Prag, Çek Cumhuriyeti (2007)

[15] Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryiğit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: A language-independent system for data-driven dependency parsing. Natural Language Engineering Journal **13**(2), 99–135 (2007)

[16] Oflazer, K.: Two-level description of Turkish morphology. Literary and Linguistic Computing **9**(2), 137–148 (1994)

[17] Oflazer, K.: Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. Computational Linguistics **22**(1), 73–90 (1996)

[18] Oflazer, K.: Statistical machine translation into a morphologically complex language. Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Sayfa 376–387 (2008)

[19] Oflazer, K., Inkelas, S.: The architecture and the implementation of a finite state pronunciation lexicon for Turkish. Computer Speech and Language **20**(1) (2006)

[20] Oflazer, K., Kuruöz, İ.: Tagging and morphological disambiguation of Turkish text. Proceedings of the Fourth Conference on Applied Natural Language Processing, Sayfa 144–149. Association for Computational Linguistics, Stuttgart, Almanya (1994)

[21] Oflazer, K., Say, B., Hakkani-Tür, D.Z., Tür, G.: Building a Turkish treebank. A. Abeillé (ed.) Treebanks: Building and Using Parsed Corpora, Sayfa 261–277. Kluwer, Londra (2003)

[22] Oflazer, K., Tür, G.: Combining hand-crafted rules and unsupervised learning in constraint-based morphological disambiguation. E. Brill, K. Church (eds.) Proceedings of the ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing (1996)

[23] Sak, H., Güngör, T., Saraçlar, M.: Morphological disambiguation of Turkish text with perceptron algorithm. CICLing 2007, vol. LNCS 4394, Sayfa 107–118 (2007)

[24] Sak, H., Güngör, T., Saraşlar, M.: Resources for Turkish morphological processing. Language Resources and Evaluation **45**(2), 249–261 (2011)

[25] Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M.: Balkanet: A multilingual semantic network for Balkan languages. Proceedings of the 1st Global Wordnet Conference. Mysore, Hindistan (2002)

[26] Wickwire, D.E.: The "sevmek thesis", a grammatical analysis of the Turkish verb system illustrated by the verb "sevmek"-to love. Master Tezi, Pacific Western Üniversitesi (1987)

[27] Yeniterzi, R., Oflazer, K.: Syntax-to-morphology mapping in factored phrase-based statistical machine translation from English to Turkish. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Sayfa 454–464. Association for Computational Linguistics, Uppsala, İsveç (2010)

[28] Yuret, D., Türe, F.: Learning morphological disambiguation rules for Turkish. Proceedings of HLT/NAACL-2006, Sayfa 328–334. New York, ABD (2006)