itüdergisi/***d***
mühendislik
Cilt:7, Sayı:4, 106-117
Ağustos 2008

Türkçe'nin olasılık tabanlı bağlılık ayrıştırması

Gülşen ERYİĞİT[*1], Eşref ADALI[1], Kemal OFLAZER[2]

[1]İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Programı, 34469, Ayazağa, İstanbul
[2]Sabancı Üniversitesi, Bilgisayar Mühendisliği Programı, 34956, Tuzla, İstanbul

Özet

Bu çalışma, Türkçe için geliştirilmiş ilk istatistiksel bağlılık ayrıştırıcısının sonuçlarını sunmaktadır. Türkçe, tümce içi öğe dizilişleri serbest, karmaşık bir çekimsel ve türetimsel biçimbirime sahip olan bitişken bir dildir ve bu özellikleri ile istatistiksel ayrıştırma konusunda ilginç sorunlar ortaya koymaktadır. Türkçe'de, bağlılık ilişkileri "çekim kümesi" adı verilen sözcük parçacıkları arasında kurulmaktadır. Bu bağlılıkların bulunması amacı ile Türkçe'nin karmaşık yapısının ayrıştırma sırasında nasıl modelleneceğinin irdelenmesi gerekmektedir. Bu çalışmada, ayrıştırma için farklı gösterim birimleri kullanan olasılık tabanlı modeller incelenmiştir. Başlangıç olarak biri kural tabanlı bir ayrıştırıcı olmak üzere üç dayanak model geliştirilmiştir. Gerçekleştirilen üç olasılık tabanlı modelin, dayanak modellere ve birbirlerine oranla başarımları değerlendirilmiştir. Ayrıştırıcının eğitimi ve sınaması için Odtü Sabancı Türkçe ağaç yapılı derlemi kullanılmıştır. Çalışma ayrıca bu derlem üzerinde sınanmış ve sonuçlaı raporlanmış ilk çalışmadır. Bu ilk incelemede, derlemin sadece sağa bağımlı (iye sözcüklerin uydu sözcüklerin sağ taraflarında yer aldığı) türde ve kesişmeyen bağlılıklar içeren bir alt kümesini ayrıştırmaya odaklanılmıştır. Eldeki derlemin boyutu nedeni ile görünüm bilgisi (sözcüğün tümünün veya gövdesinin ayrıştırma birimi gösterimlerinde bir özellik olarak kullanılması) kullanmayan ve sadece birimler arası etiketsiz bağlılıkları bulmaya yönelik incelemeler yapılmıştır. Sonuçlarımız, çekim kümeleri arasındaki doğru bağlıkların bulunma başarımı gözönüne alındığında, ayrıştırma birimi olarak çekim kümelerinin kullanıldığı ve bağlam bilgisinden yararlanan modelin en yüksek başarımı sağladığını göstermektedir.

Anahtar Kelimeler: *Bağlılık ayrıştırması, doğal dil işleme, ayrıştırma, sentaks analizi.*

[*]Yazışmaların yapılacağı yazar: Gülşen Eryiğit. gulsen.cebiroglu@itu.edu.tr; Tel: (212) 285 67 03.
Bu makale, birinci yazar tarafından İTÜ Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Programında tamamlanmış olan "Türkçe'nin bağlılık ayrıştırması" adlı doktora tezinden hazırlanmıştır. Makale metni 28.12.2006 tarihinde dergiye ulaşmış, 22.01.2007 tarihinde basım kararı alınmıştır. Makale ile ilgili tartışmalar 01.03.2009 tarihine kadar dergiye gönderilmelidir.

Probabilistic dependency parsing of Turkish

Extended abstract

This paper presents results from the first statistical dependency parser for Turkish. Turkish is a free-constituent order language with complex agglutinative inflectional and derivational morphology and presents interesting challenges for statistical parsing, as in general, dependency relations are between "portions" of words – called inflectional groups. We have explored statistical models that use different representational units for parsing. We have used the Turkish Dependency Treebank to train and test our parser but have limited this initial exploration to that subset of the treebank sentences with only left-to-right non-crossing dependency links. Our results indicate that the best accuracy in terms of the dependency relations between inflectional groups is obtained when we use inflectional groups as units in parsing, and when contexts around the dependent are employed.

Turkish shows very different characteristics from the well-studied languages in parsing literature. Many of these characteristics are common for all agglutinative languages such as Basque, Estonian, Finnish, Hungarian, Japanese and Korean. It is a flexible constituent order language. Even though in written texts, the constituent order of sentences generally conforms to the SOV or OSV structures, the constituents may freely change their position depending on the requirements of the discourse context. From the point of view of dependency structure, Turkish is predominantly (but not exclusively) head final. Furthermore, Turkish morphotactics is quite complicated: a given word form may involve multiple derivations and the number of word forms one can generate from a nominal or verbal root is theoretically infinite. Derivations in Turkish are very productive, and the syntactic relations that a word is involved in as a dependent or head element, are determined by the inflectional properties of the one or more (possibly intermediate) derived forms. In this work, we assume that a Turkish word is represented as a sequence of inflectional groups (IGs hereafter), separated by ˆDBs, denoting derivation boundaries. A sentence would then be represented as a sequence of the IGs making up the words. When a word is considered as a sequence of IGs, linguistically, the last IG of a word determines its role as a dependent, so, syntactic relation links only emanate from the last IG of a (dependent) word, and land on one of the IGs of a (head) word on the right (with minor exceptions). And again with minor exceptions, the dependency links between the IGs, when drawn above the IG sequence, do not cross.

We implemented three baseline parsers:
1. The first baseline parser links a word-final IG to the first IG of the next word on the right.
2. The second baseline parser links a word-final IG to the last IG of the next word on the right.
3. The third baseline parser is a deterministic rule-based parser that links each word-final IG to an IG on the right based on the approach of Nivre (2003). The parser uses 23 unlexicalized linking rules and a heuristic that links any non-punctuation word not linked by the parser to the last IG of the last word as a dependent.

In addition to these, we implemented three probabilistic models:
1. "Unlexicalized" Word-based Model, where the words are represented as the concatenation of their IGs and are used as the parsing unit during the parsing.
2. IG-based Model, where each word is splitted into its IGs and then the IGs are used as the smallest parsing unit.
3. IG-based Model with Word-final IG Contexts, where the IGs are again used as the parsing unit. This model differs from the previous one in the way it uses the contextual units and calculates the distances between units.

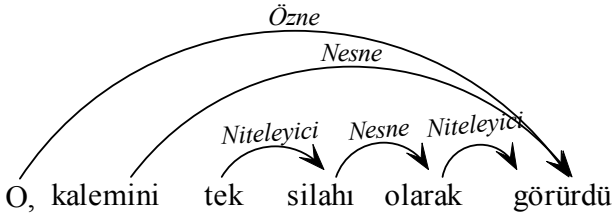
Our results indicate that all of our models perform better than the three baseline parsers, even when no contexts around the dependent and head units are used. We get our best results with Model 3, where IGs are used as units for parsing and contexts are comprised of word final IGs. The highest accuracy in terms of percent of correctly extracted IG-to-IG relations excluding punctuations (73.5%) was obtained when one word is used as context on both sides of the the dependent. We also noted that using a smaller treebank to train our models did not result in a significant reduction in our accuracy indicating that the unlexicalized models are quite effective, but this also may hint that a larger treebank with unlexicalized modeling may not be useful for improving link accuracy.

Keywords: *Dependency parsing, natural language processing, parsing, syntax analysis*

Giriş

Farklı türlerde oluşturulmuş ağaç yapılı derlemlerin sayısının artması ile birlikte, bu derlemler üzerinde eğitilen istatistiksel ayrıştırıcıların geliştirilmesi de hız kazanmıştır. Sözcükler arası bağlılıkların ayrıştırmadaki önemli etkisinin anlaşılması (Charniak, 2000; Collins, 1996) ile birlikte bağlılık gramerleri ayrıştırma alanında sıkça kullanılır hale gelmişlerdir. Bunlara örnek olarak İngilizce için Yamada ve Matsumoto (2003)'nun, Japonca için Kudo ve Matsumoto (2000; 2002)'nun, Sekine vd. (2000)'nin, Korece için Chung ve Rim (2004)'in, İsveççe için Nivre vd. (2004)'nin, Çekçe için Nivre ve Nilsson (2005)'ın çalışmaları gösterilebilir.

Bağlılık gramerleri, tümcenin yapısını sözcükler arasında ikili bağlılık ilişkileri kurarak gösterirler. Örneğin Şekil 1 Türkçe bir tümcenin bağlılık grafiğini göstermektedir. Bu grafikte, bağlılık etiketleri uydu sözcükten iye sözcüğe doğru çizilen okların üzerinde belirtilmektedir.



Şekil 1. Bağlılık grafiği

CFG alt yapısı kullanan ayrıştırıcılar, öğelerin tümce içerisinde tümcenin genel anlamını bozmadan serbestçe yer değiştirebildiği dillerde daha az elverişli görülmektedirler. Collins vd. (1999), Collins (1997)'in İngilizce için geliştirdiği ayrıştırıcısını Çekçe'ye uyarlamış ve İngilizce için elde edilen sonuçlara oranla başarımda önemli ölçüde düşüş olduğunu gözlemlemişlerdir.

Türkçe

Bitişken bir dil olan Türkçe'de, sözcüklerin sonlarına ard arda çekim ve türetim ekleri konularak yüzlerce farklı sözcük oluşturulabilir. Tümcelerin, sözcük dizilişleri itibari ile büyük çoğunlukla ÖNY (Özne-Nesne-Yüklem) veya NÖY kalıbına uymasına rağmen, öğeler anlatılmak istenen içeriğe ve vurguya bağlı olarak tümce içerisinde serbestçe yer değiştirebilirler.

Türkçe'nin biçimbirimi oldukça karmaşık bir yapıya sahiptir: bir sözcük içerisinde, birden çok türeme görülebilir ve bir isim veya eylem kökünden üretilebilecek farklı sözcüklerin sayısı kuramsal olarak sonsuzdur. Türkçe'nin türetim sistemi çok üretkendir ve bir sözcüğün uydu veya iye olarak içerisinde bulunduğu tümce ilişkileri, sözcüğün içerdiği bir veya daha fazla türemiş yapının biçimbirimsel özellikleriyle belirlenmektedir.

Bu çalışmada, Türkçe bir sözcüğün bir dizi çekim kümesinden (ÇK) oluştuğu ve bu ÇK'lerin türetim sınırlarından (TS) bölündüğü varsayılmaktadır. Her çekim kümesi ilgili bölüme ait biçimbirimsel bilgiyi barındırır:

gövde+$ÇK_1$+^TS+$ÇK_2$+^TS +....+^TS+$ÇK_n$

Burada her $ÇK_i$ ilgili biçimbirimsel özellikleri ve sözcük sınıflarını belirtmektedir. Örnek olarak, türemiş bir niteleyici olan "sağlamlaştırdığımızdaki" sözcüğü şu şekilde gösterilmektedir[1]:

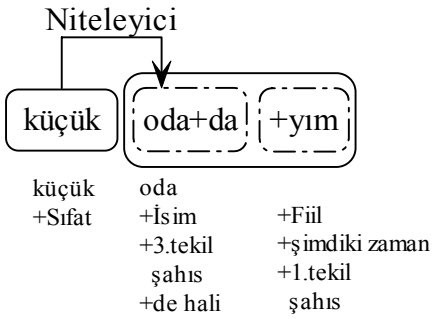
sağlam+Adj
+ˆTS+Verb+Become
+ˆTS+Verb+Caus+Pos
+ˆTS+Noun+PastPart+A3sg+P3sg+Loc
+ˆTS+Adj+Rel

Buradaki beş ÇK, TS işaretleri ile birbirinden ayrılmış özellik dizileridir. İlk ÇK gövdenin tek özelliği olan sözcük sınıfını göstermektedir. İkinci ÇK, önceki sıfata "oluşmak" anlamı katılarak bir eylem türetmeyi göstermektedir. Üçüncü ÇK önceki eylemden olumlu bir ettirgen eylemin türetildiğini belirtmektedir. Dör-

[1] Derlem gösterimi ile belirtilmiş biçimbirimsel özellikler ve kategoriler şöyledir: +Adj: Sıfat, +Verb: Eylem, +Become: oluşmak, +Caus: Ettirgen, +Pos: olumlu, +Noun: İsim, +PastPart: geçmiş zaman ortacı, +A3sg: 3. tekil kişi kişi/sayı uyum imi, +P3sg: 3. tekil kişi iyelik imi, +Loc: -de hali, +Rel: ilişkilendirici

düncü ÇK, alt sözcük sınıfı olarak geçmiş zaman ortacı ve bunun yanında bazı biçimbirimsel özellikler taşıyan bir ismin türetilmesini belirtmektedir. Son olarak da beşinci ÇK, ilişkilendirici bir sıfat türetilmesini belirtmektedir.

Bağlılıkları sadece sözcükler arasında göstermek ayrıştırma işlemi için yeterince anlamlı bilgi taşımamaktadır. Şekil 2'de *"küçük odadayım"* tümcesi içerisindeki bağlılık gösterilmektedir. Sözcüklerin kökleri ve biçimbirimsel çözümlemeleri alt taraflarında verilmektedir. Örnekte *küçük* olan, *odadayım* sözcüğü değil *oda*'dır. *odadayım* isimden eyleme dönüşmüş bir sözcüktür. İki sözcük arasında kurulan bağlantı *odadayım* sözcüğünün eyleme dönüşmeden önceki isim halinden kaynaklanmaktadır. Bu durum, sıfatların genel olarak isimlere bağlanması kuralından kaynaklanmaktadır.



Şekil 2. Çekim kümesi yapıları

Buradan yola çıkarak, ayrıştırıcının bulduğu bağlılıklar, sadece uydu ve iye sözcüğü değil, uydu ve iye ÇK'yi de belirtmelidir. Bağlılıklar uydu sözcüğün sadece son ÇK'sinden çıktıkları için, bağlılık çıkmayan bazı ÇK'ler olacaktır (örn., Şekil 3 (Oflazer, 2003)'deki *büyümesi* sözcüğünün ilk ÇK'si). Bu tip ÇK'lerin aynı sözcük içerisindeki hemen sağ taraflarında yer alan ÇK'ye bağlandıkları varsayılmaktadır. Ancak bu tür ilişkileri tespit etmek biçimbirimsel çözümleyicinin görevi olduğundan, ayrıştırıcı bunları belirlemeye çalışmaz. İleriki bölümlerde anlatılacak olan ayrıştırma modellerinin tümü, ÇK'ler arasındaki sözcük içi olmayan bağlılık ilişkilerini bulmayı amaçlamaktadırlar. Buna ek olarak, kullanılacak başarım ölçütleri, ÇK'lere ve bunlar arasındaki ilişkilerin belirlenmesine dayanmaktadır.

Tümce yapısı modeli olarak Şekil 4'deki yapı kullanılmaktadır. Şekilde, üst taraf tümce içerisindeki sözcükleri göstermektedir. Biçimbirimsel çözümlemeden ve belirsizlik gideriminden sonra her sözcük içerdiği çekim kümelerinin sıralanması ile gösterilir. Daha sonra, çekim kümeleri yeniden numaralandırılarak ayrıştırma sırasında bir "birim" olarak kullanılırlar. "*" ile işaretlenen çekim kümeleri bağlılık oklarının çıkabileceği çekim kümelerini göstermektedir. Bu tür çekim kümelerinin sayısı tümce içerisindeki toplam sözcük sayısına eşittir. Tümcenin genelinin "iye"si olan ÇK dışında (bu ÇK hiçbir yere bağlanmayacaktır), bu tür ÇK'lerin hepsinden bir bağlılık oku çıkacaktır.

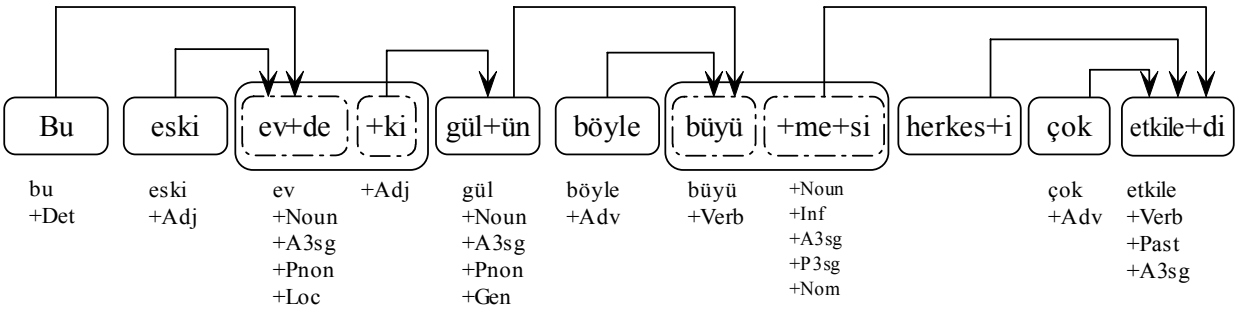
Makalenin geri kalanında, olasılık tabanlı bağlılık ayrıştırması yönteminin genel bir anlatımı yapılacak ve daha sonra Türkçe'nin bağlılık ayrıştırması için geliştirilmiş üç model tanıtılacaktır. Bundan sonra, bu modeller ile elde edilen başarımlar ve en iyi model üzerinde yapılan bazı deneyler sunulacaktır. Makale, sonuçların tartışılması, ayrıştırıcının yaptığı hataların incelenmesi ve sonuç bölümü ile son bulacaktır.

Ayrıştırıcı

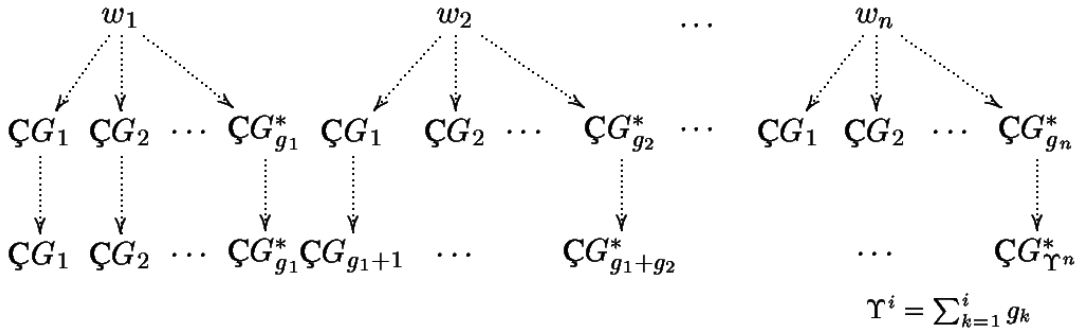
Olasılık tabanlı bağlılık ayrıştırıcıları öncelikle birimler arası bağlılıkların olasılıklarını bulur ve daha sonra arama uzayındaki olası bütün bağlılık ağaçları içerisindeki en olası bağlılık ağacını (T*) bulmaya çalışırlar:

$$T^* = \arg\max_T P(T \mid S) = \arg\max_T \prod_{i=1}^{n-1} P(bag(u_i, u_{H(i)}) \mid S) \quad (1)$$

Bu denklemde *S*, *n* adet birimin sıralanmasıyla oluşan tümcedir. $u_{H(i)}$, uydu u_i biriminin bağlandığı iye birimi ifade etmektedir. *T* ise *bag* $(u_i, u_{H(i)})$ uydu-iye bağlılık ilişkilerini içeren olası tüm bağlılık ağaçlarını ifade etmektedir.



Şekil 3. Türkçe bir tümcedeki bağlılıklar



Şekil 4. Tümce yapısı

Türkçe derlemdeki bağlılıkların %95'inin sağa bağımlı türde bağlılıklar olmasından yola çıkarak, bu ayrıştırıcıda ayrıştırma algoritması olarak Sekine vd. (2000)'nin "*Geriye doğru demetli arama*" algoritması kullanılmıştır. İlk olarak sağa bağımlı türde bir dil olan Japonca için geliştirilen bu algoritma, tümceyi sondan başlayarak başa doğru ayrıştırır. Algoritma her adımda üzerinde bulunduğu birimi, tümce içerisinde bu birimin sağ tarafında yer alan birimlerden birine bağlamaya çalışır. Bu işlem sırasında, ayrıştırması kısmen yapılmış tümce üzerinde, en olası *d* adet kısmi ayrıştırmayı bir demet içerisinde tutar. Her adımda, ortaya çıkan yeni ayrıştırmalar içerisinde, demettekilerden daha yüksek olasılıklı olan ayrıştırmalar, demette yer alan daha düşük olasılıklı ayrıştırmaların yerine geçerler. Bu algoritma, aynı zamanda, kesişmeyen bağlılık ilkesini benimsemektedir. Bu nedenle, uydu birimlerin bağlanacakları iye birimler seçilirken yeni kurulan bağlılıkların daha önceden oluşmuş kısmi bağlılıklarla çakışmamasına dikkat edilir.

Geriye doğru demetli arama algoritması kısmi olarak oluşturduğu yapıların olasılıklarını hesaplamak için Denklem 1'i kullanır. *P(bag(u$_i$,u$_{H(i)}$)|S)* ile belirtilen ikili bağlılık olasılıklarını hesaplamak olasılık tabanlı ayrıştırma modelinin görevidir. Bu ayrıştırıcıda, Collins (1996)'in olasılık tabanlı modelinin bir çeşidi olan Chung ve Rim (2004)'in yaklaşımı benimsenmiştir. Denklem 2, kullanılan modelin denklemini vermektedir. Bu denklemde *P(bag,(u$_i$,u$_{H(i)}$)|S)* uydu u$_i$'nin bir *u$_{H(i)}$* iye birimine bağlanma olasılığıdır ve iki diğer olasılığın hesaplanması ile bulunur:

$$P(bag(u_i, u_{H(i)}) \mid S) \approx P(ilk(u_i, u_{H(i)}) \mid \Phi_i \Phi_{H(i)}) \cdot P(u_i\text{'nin uzk}(i, H(i)) \text{ uzağa bağlanması} \mid \Phi_i) \quad (2)$$

P(ilk(u$_i$,u$_{H(i)}$)|Φ$_i$ Φ$_{H(i)}$) (Benzer bir bağlılığın benzer bir bağlam içerisinde görülme olasılığı. *Φ$_i$*, *u$_i$* uydu birim etrafındaki bağlam bilgisini ve

$\Phi_{H(i)}$, $u_{H(i)}$ iye birim etrafındaki bağlam bilgisini belirtmektedir.)

P(u_i'nin uzk(i,H(i)) uzağa bağlanması| Φ_i) (Uydu birimin benzer bağlamda benzer uzaklıktaki bir iye birime bağlanma olasılığı. Uydu ve iye birim arasındaki uzaklık bir uzaklık fonksiyonu kullanılarak hesaplanır.)

Aşağıda anlatılacak olan ayrıştırma modelleri için kullanılacak istatistiksel parametreler Türkçe ağaç yapılı derlemden (Oflazer vd., 2003) hesaplanmıştır. Derlem, diğer diller için hazılanmış derlemlere (örn., Penn Treebank) oranla daha küçük boyutta olduğu için, özellikle Türkçe için sıkça karşılaşılan veri seyrekliği sorununu azaltabilmek, ikili bağlanma olasılıkları hesaplanırken görünüm bilgisi kullanmayan (bir diğer deyişle sadece bazı biçibirimsel ve sentaks özelliklerini dikkate alan) bir yol izlenmiştir. Bu seçim yapılırken aynı zamanda, görünüm bilgisi kullanmayan ayrıştırıcıların (Klein ve Manning, 2003; Chung ve Rim, 2004) başarımından esinlenilmiştir.

Ayrıştırma modellerinin ayrıntıları

Bu bölümde, Türkçe için sınanan üç modelin ayrıntıları verilmektedir. Bunların her üçü de görünüm bilgisi kullanmayan modellerdir ve birbirlerinden ayrıştırma birimlerini nasıl kullandıkları ve bağlamı nasıl modelledikleri yönünden farklılık gösterirler. Her üç modelde de, Denklem 2'deki olasılık modeli kullanılmıştır.

ÇK etiketlerinin indirgenmesi

Derlem oluşturulurken sözcüklerin analizi için kullanılan biçimbirimsel çözümleyici (Oflazer, 1994), sentaks ve anlamsal bilgi de dahil olmak üzere oldukça zengin bir çözümleme bilgisi sunmaktadır. Ancak bu bilgilerin tümü ayrıştırma için gerekli değildir. Buna ek olarak, bu bilgilerin farklı alt kümeleri sözcüğün görevine bağlı olarak gerekli olabilir. Aşağıda anlatılacak modellerde, ÇK'lerin gösteriminde şu indirgeme yöntemi kullanılmıştır:

▪ ÇK bir uydu olarak kullanıldığında,

- Eğer isim türünden[2] bir ÇK ise, o zaman sadece durum imi ile belirtilir.
- Diğer türden ÇK'ler, sadece ana sözcük sınıfları ile belirtilirler.

▪ ÇK bir iye olarak kullanıldığında,

- Eğer isim türünden bir ÇK veya zaman ortacı olan bir sıfat[3] ÇK ise, o zaman ana sözcük sınıfı ve iyelik uyum imi ile birlikte ifade edilir.
- Diğer türden ÇK'ler, sadece ana sözcük sınıfları ile belirtilirler.

Bu tür indirgenmiş bir gösterim, bir çok farklı sözcük türünün sadece ortak özelliklerini birleştirerek istatistik toplanmasını sağladığından dolayı, aynı zamanda seyrek veri sorununu da azaltmaya yardım etmektedir.

Denklem 2'de sağ taraftaki (uydu ve iye arasındaki uzaklık ile ilgili olan) olasılık hesaplanırken, öncelikle derlem tümceleri üzerinden istatistikler çekilmiştir. Ayrıştırma birimi olarak sözcükler alındığında, bağlılıkların %90'ının 3 veya daha yakın uzaklıkta bir sözcükte sonlandığı görülmüştür. Benzer şekilde, ayrıştırma birimi olarak ÇK'ler alındığında, bağlılıkların %90'ının 4 veya daha yakın uzaklıkta bir ÇK'de sonlandığı görülmüştür. Bu nedenle, uzaklığın sözcük temelinde hesaplandığı model 1 ve model 3'de uzaklık fonksiyonu eşik değeri olarak *k* parametresi 4, uzaklığın ÇK temelinde hesaplandığı ÇK tabanlı model 1'de ise k=5 olarak alınmıştır. Eşik değerinden yüksek olan uzaklıklar eşik değerine çekilerek hesaplanmışlardır. Testlerde demet boyu için en yüksek başarımı veren d=3 değeri kullanılmıştır.

Düzleştirme algoritması olarak Collins (1996)'in çalışmasında kullanılan "*düşürerek düzleştirme"* algoritmasının benzer bir yöntem kullanıl-

[2] Bunlar isimler, zamirler ve mastar, zaman ortacı almış türemiş halleridir. Sadece isim türünden ÇK'ler durum imine sahiptirler ve bunların uydu olarak rolünü belirleyen imler esas olarak durum imleridir.

[3] Şimdiki/Geçmiş/Gelecek zaman ortacına sahip sıfatlar, isim türünden ÇK'ler dışında iyelik uyum imine sahip tek ÇK tipleridir.

mıştır. Denklem 2, iye ve uydu birimin bağlam bilgilerinin hepsinin birden bir seferde kaldırılması ile elde edilen olasılık değerleri ile aradeğerlenerek hesaplanmıştır.[4] Buna göre, asıl yürütmeler sırasında $P(ilk(u_i,u_{H(i)})|\Phi_i\ \Phi_{H(i)})$ düzleştirilmiş olasılığı derlemden çıkarılmış iki düzleştirilmemiş olasılık aradeğerlenerek hesaplanmıştır:

$P_1(ilk(u_i,u_{H(i)})|\Phi_i\ \Phi_{H(i)})$ *ve* $P_2(ilk(u_i,u_{H(i)}))$.

Eğer bu aradeğerlendirmeden sonra bile olasılık değeri sıfır çıkıyorsa o zaman olasılık değeri olarak sıfıra yakın çok küçük bir değer atanmıştır. Benzer aradeğerleme, denklemin ikinci kısmı için de yapılmıştır.

Model 1 – Sözcük tabanlı model
Bu modelde, bir sözcük uydu olarak kullanıldığında son ÇK[5]'sinin indirgenmiş gösterimi ile, iye olarak kullanıldığında ise tüm ÇK'lerinin indirgenmiş gösterimlerinin birleşimi ile ifade edilir. Bir sözcük hem iye hem de uydu olabileceğinden, kullanılacak indirgenmiş gösterim ayrıştırma sırasında dinamik olarak belirlenecektir.

Bu aşamadan sonra ayrıştırma, birim olarak sözcükleri kullanarak devam eder. Ayrıştırıcı bu birimler arasında bağlılıkları kurduktan sonra, bağlılıklar gerçek ÇK-ÇK bağlılıklarını belirlemek üzere ÇK'lere eşleştirilir. Bir uydudan çıkan bağlılığın son ÇK'den çıktığı bilinmektedir. İye sözcük için, bağlılığın sözcüğün ilk ÇK'sinde sonlandığı varsayılır.[6]

Bağlam içerisinde yer alan komşu birimlerin gösterimleri de ilişkili oldukları birime göre olan konumlarına bağlı olarak belirlenecekler-dir. İlişkili olduğu birimin sol tarafında yer alan komşu birimler uydu olarak (son ÇK'si ile), sağ tarafında yer alanlar ise iye olarak (ÇK'lerinin birleşimi ile) işlem göreceklerdir. Buradaki ve sonraki modellerde, komşuların üst üste gelme durumları gözardı edilmektedir.

Şekil 5, Şekil 3'de verilen örnek tümce üzerinde biçimbirimsel çözümlemeleri ve her üç modelde kullanılan indirgenmiş etiketleri göstermektedir. Her model için, birimin uydu olarak ve iye olarak kullanıldığında etiketinin ne olacağı listelenmiştir. Model 1 için 3. ve 4. satırlardaki etiketler kullanılmaktadır.

Model 2 – ÇK tabanlı model
Bu modelde, her ÇK yukarıda anlatıldığı şekilde indirgenmiş gösterimi kullanılarak belirtilir ancak sözcüğü oluşturmak üzere biraraya getirilmezler. Dolayısıyla, ayrıştırma birimlerimiz ÇK'lerdir. Ayrıştırıcı sözcük sonu ÇK'lerinden sağ tarafta yer alan bir ÇK'ye doğrudan ÇK-ÇK bağlılığı kurar. Kullanılan bağlam, uydunun ve iyenin sol tarafında (önceki sözcüğün son ÇK'sinden başlanarak) ve sağ tarafında yer alan ÇK'lerdir.

Bu modelde kullanılan etiketler, Şekil 5'in 5. ve 6. satırlarında gösterilmektedir. 6. satırda bulunan boş hücreler sözcük sonu ÇK'leri olmadıkları için uydu olamayacak ÇK'lere aittir.

Model 3 – ÇK tabanlı model (bağlam için sözcük sonu ÇK'leri)
Bu model, model 2 ile yaklaşık olarak aynıdır. Modeller arasındaki iki fark şunlardır:

(i) bağlam için sözcük içerindeki sözcük sonu ÇK'si olmayan ÇK'ler gözardı edilerek, sadece soldaki ve sağdaki sözcük sonu ÇK'leri kullanılır (bağlamın iye ile üst üste gelmesi durumu hariç; bu durumda son ÇK yerine iye ÇK'nin etiketi kullanılır). (ii) uzaklık fonksiyonu sözcük temelinde hesaplanmıştır. Bunun nedeni uyduların tümce görevlerinin son ÇK'ler tarafından belirlenmesidir.

[4] Araştırmalar sırasında, bağlam bilgisini teker teker azaltmak veya çekim özelliklerini azaltmak gibi birçok farklı düşürerek düzleme modeli denenmiştir. Deneyler sonucunda, burada tanıtılan modelin en yüksek başarımı sağladığı gözlemlenmiştir.

[5] Sözcüğün birden çok ÇK'si varsa, son ÇK dışındakilerin bu sözcüğün iyesine nasıl bağlanacağı ile ilgili bilgi taşımadığı hatırlanmalıdır.

[6] Bu seçim, derlemdeki bağlılıkların %85.6'sının iye sözcüğün ilk (büyük olasılıkla tek) ÇK'sinde, %14.4'ünün ise ilk ÇK'den farklı bir ÇK'de sonlanmasına dayanmaktadır.

Cümle	Bu	eski	evdeki		gülün	böyle	büyümesi		herkesi	çok	etkiledi
Biçim-birimsel Çözüm-leme	bu +Det	eski +Adj	ev +Noun +A3sg +Pnon +Loc	+Adj	gül +Noun +A3sg +Pnon +Gen	böyle +Adv	büyü +Verb	+Noun +Inf +A3sg +P3sg +Nom	herkes +Pron +A3pl +Pnon +Acc	çok +Adv	etkile +Verb +Past +A3sg
Model 1 Sahip etiket	<+Det>	<+Adj>	<+Noun+Pnon+Adj>		<+Noun+Pnon>	<+Adv>	<+Verb+Noun+P3sg>		<+Pron+Pnon>	<+Adv>	<+Verb>
Model 1 bağımlı etiket	<+Det>	<+Adj>	<+Adj>		<+Gen>	<+Adv>	<+Nom>		<+Acc>	<+Adv>	<+Verb>
Model 2/3 sahip etiket	<+Det>	<+Adj>	<+Noun+Pnon>	<+Adj>	<+Noun+Pnon>	<+Adv>	<+Verb>	<+Noun+P3sg>	<+Pron+Pnon>	<+Adv>	<+Verb>
Model 2/3 bağımlı etiket	<+Det>	<+Adj>		<+Adj>	<+Gen>	<+Adv>		<+Nom>	<+Acc>	<+Adv>	<+Verb>

Şekil 5. Ayrıştırma modellerinde kullanılan etiketler

Deneysel çalışma sonuçları

Bu çalışmada, sadece sağa bağımlı[7] ve kesişmeyen[8] bağlılıkların ayrıştırılması incelendiği için, deneyler Türkçe derlemde bu kısıta uymayan bağlılıkları içeren tüm tümceler elenerek, geri kalan 3398 tümceden oluşan bir altküme üzerinde yapılmıştır. Deneylerde yetkin etiketler kullanılmıştır. Derlemdeki tümceler ortalama 8 sözcük[9] olmak üzere 2 ile 40 arasında sözcük içermektedirler; tümcelerin %90'ı 15 veya daha az sözcükten oluşmaktadırlar. ÇK temelinde incelendiğinde, 10 ÇK ortalamayla, tümceler 2 ile 50 arasında ÇK içerirler; tümcelerin %90'ı 15 veya daha az ÇK'den oluşmaktadır. Veri kümesi, 10 katlı çapraz doğrulama için on farklı şekilde, eğitim ve sınama kümelerine bölünmüştür. Geliştirdiğimiz modellerin başarımlarına bir taban oluşturmak üzere üç dayanak model geliştirilmiştir:

1. Dayanak 1: her sözcük (son ÇK'sinden) sağındaki sözcüğün ilk ÇK'sine bağlanır.
2. Dayanak 2: her sözcük (son ÇK'sinden) sağındaki sözcüğün son ÇK'sine bağlanır.[10]
3. Dayanak 3: gerekirci kural tabanlı bir ayrıştırıcıdır. Nivre (2003)'nin ayrıştırma algoritmasına benzer bir algoritma ve elle hazırlanmış 23 adet görünüm bilgisi içermeyen kural kullanılarak sözcükler son ÇK'lerinden sağ taraflarındaki bir ÇK'ye bağlanırlar. Ayrıştırma sonucunda, ayrıştırıcı tarafından bağlanmadan kalan, noktalama işareti olmayan sözcükler son sözcüğe uydu olarak bağlanırlar.

Tablo 1 dayanak ve olasılık tabanlı modeller ile elde edilen deney sonuçlarını göstermektedir. Tanıtılan üç olasılık tabanlı model uydunun ve

[7] Derlemdeki bağlılıkların %95'i sağa bağımlı türde bağlılıklardan oluşmaktadır.

[8] Kesişmeyen: kesişen veya uydu-iye bağlılık oku altında iye sözcükten bağımsız herhangi bir sözcük barındırmayan. Derlemdeki bağlılıkların %2.5'u başka bir bağlılığı kesmektedir.

[9] Bu sayının diğer dillere göre az olması oldukça normaldir. Diğer dillerde ayrı yazılan işlevsel sözcükler Türkçe'de ekler vasıtası ile diğer sözcüklerle bir arada yazılırlar.

[10] Tek ÇK'den oluşan iye sözcükler için Dayanak Model 1 ve 2 aynı şekilde davranırlar.

iyenin etrafında farklı bağlamlarla sınanmışlardır. Sütun 3 ve 4, noktalama işaretleri dahil ve hariç olmak üzere tüm birimler için, ÇK'ler arası doğru bulunan bağlılık yüzdesini vermektedir. Sütun 5 ve 6, tüm bağlılıkları derlemdeki ile aynı olarak bulunmuş tümcelerin yüzdesini vermektedir. Tablodaki her değer 10 katlı çapraz doğrulama sırasında gerçekleştirilen 10 çalıştırma sonucunda elde edilen başarımların ortalamasını ve standart hatasını ifade etmektedir. *Amacımız 4. sütunda verilen doğru olarak belirlenmiş ÇK-ÇK bağlılık yüzdesi için en yüksek değeri elde etmektir.* Bu deneylerde en iyi başarımlar, model 3 ile, bağlam olarak uydu birimin her iki tarafından da birer birim kullanılması durumunda elde edilmiştir. Aynı bağlam boyutunda model 2'den az bir farkla iyi olmasına karşın, her 10 iterasyon için ortalamalar arasındaki fark (0.4 ± 0.2) istatistiksel olarak belirgindir.

Görünüm bilgisi içermeyen modeller kullandıldığı için, daha küçük bir eğitim verisi kullanmanın, modellerin başarımına büyük bir etkisi olup olmadığının görülmesi amacıyla Tablo 2'de sonuçları verilen deneyler hazırlanmıştır. Tablo, model 3 (bağlamsız ve uydunun etrafından birer birim kullanılarak) için derlemin 1500 tümcesi kullanılarak yapılan deneylerin sonuçlarını vermektedir. Eğitim kümesinin boyutunun sonuçlar üzerinde küçük bir etkisi olduğu görülmektedir.

Bu makalede, Türkçe için geliştirilmiş bir ayrıştırıcının başarımı ölçülürken sadece sözcükler arasındaki bağlılık başarımını ölçmenin doğru bir yaklaşım olmadığı belirtilmiştir. Ancak, diğer sözcük tabanlı yaklaşımlar için, karşılaştırma yapılabilmesine yönelik olarak sözcük-sözcük başarım değerlendirmesi de yapılmıştır. Bu değerlendirmede, bir uydu sözcük için iye sözcük (doğru iye ÇK olup olmadığına bakılmaksızın) doğru olarak saptandığında bağlılığın doğru olduğu varsayılmıştır. Tablo 3, Tablo 1'deki en iyi başarımları veren parametreler kullanılarak modellerin sözcük-sözcük başarımlarını vermektedir.

Yukarıda tanıtılan modellere ek olarak, ayrıştırıcımız, tamamen saf bir sözcük tabanlı modelle de sınanmıştır. Bu modelde hem uydu hem de iye sözcük ÇK'lerinin birleşimi (tüm ÇK'lerinin indirgenmiş gösterimlerinin birleşimi) ile ifade edilmiştir. Bu modelin başarımı[11] Tablo 3'ün son satırında verilmiştir. Bu sonuç kural tabanlı dayanak model 3'ün başarımdan bile daha düşüktür. Bu modelde de, model 1'de yaptığımız gibi uyduyu iye sözcüğün ilk ÇK'sine bağlarsak, noktalama işaretleri gözardı edilerek elde edilen ÇK-ÇK başarımı 69.9 ± 3.1 olmaktadır. Bu değer de yine dayanak 3'ün başarımından (70.5 ± 0.8) daha düşüktür.

Tartışma

Sonuçlar her üç modelimizin de, uydu ve iye birimler etrafında bağlam bilgisi kullanılmasa bile, dayanak modellerimizden daha iyi sonuç verdiğini belirtmektedir. En iyi sonuçlar, ayrıştırma birimi olarak ÇK'lerin ve bağlam bilgisi olarak sözcük sonu ÇK'lerinin kullanıldığı model 3 ile elde edilmiştir. Elde edilen en yüksek başarım, bağlam olarak uydu birimin sağından ve solundan birer birim kullanıldığında elde edilmiştir.[12] Ayrıca, daha küçük boyutta bir eğitim kümesi kullanılmasının keskin bir düşüşe neden olmadığı gözlemlenmiştir. Bu durum, görünüm bilgisi kullanmayan modellerin oldukça verimli olduğunu belirtmektedir. Ancak, bu sonuç daha büyük boyutta bir derlem kullanmanın bile başarımı arttırmayacağına dair bir ipucu olarak da görülebilir.

En iyi model ile elde edilen sonuçlara daha ayrıntılı bir bakış (Tablo 4), başarımın artan tümce uzunluğu ile düşüşe geçtiğini göstermektedir. Uzun tümceler için, büyük olasılıkla görünüm bilgisi de içeren daha karmaşık modellerin kullanılması gerektiği düşünülmektedir. En iyi sonuç veren modelin hataları üzerinde yapılan daha derinlemesine bir incelemeyle, hataların %40'nın doğru biçimbirimsel özelliklerde ancak yanlış konumdaki ÇK'lere bağlanılmasından kaynaklandığı görülmektedir. Bu durum uzak-

[11] Aynı zamanda, Model 1'in bağlam kullanmayan (79.1 ± 1.1) başarımından da daha düşük.

[12] Yararlı olacağına inanılan farklı biçimbirimsel özelliklerin kullanılması ile ilgili yapılan önceki deneylerde daha düşük başarımlar elde edilmiştir.

Tablo 1. Dayanak modeller ve olasılık tabanlı modeller ile ayrıştırma sonuçları

Dl=1 ve Dr=1 uydunun, Hl=1 ve Hr=1 iyenin solundan ve sağından birer birim kullanmayı ifade eder.

		ÇK-ÇK bağlılıklarının doğru olma yüzdesi		Tüm bağlılıkları doğru olan tümcelerin yüzdesi	
Model	Bağlam	Sözcük+Nokt.	Sözcük	Sözcük+Nokt.	Sözcük
Dayanak 1	-	59.9±0.3	63.9±0.7	21.4±0.6	24.0±0.7
Dayanak 2	-	58.3±0.2	62.2±0.8	20.1±0.0	22.6±0.6
Dayanak 3	-	69.6±0.2	70.5±0.8	31.7±0.7	36.6±0.8
Model 1	Yok	69.8±0.4	71.0±1.3	32.7±0.6	36.2±0.7
(k=4)	Dl=1	69.9±0.4	71.1±1.2	32.9±0.5	36.4±0.6
	Dl=1 Dr=1	71.3±0.4	72.5±1.2	33.4±0.8	36.7±0.8
	Hl=1 Hr=1	64.7±0.4	65.5±1.3	25.4±0.6	28.7±0.8
	Dl=1 Dr=1 Hl=1 Hr=1	71.4±0.4	72.6±1.1	34.2±0.7	37.2±0.6
Model 2	Yok	70.5±0.3	71.9±1.0	32.1±0.9	36.3±0.9
(k=5)	Dl=1	71.3±0.3	72.7±0.9	33.8±0.8	37.4±0.7
	Dl=1 Dr=1	71.9±0.3	73.1±0.9	34.8±0.7	38.0±0.7
	Hl=1 Hr=1	57.4±0.3	57.6±0.7	23.5±0.6	25.8±0.6
	Dl=1 Dr=1 Hl=1 Hr=1	70.9±0.3	72.2±0.9	34.2±0.8	37.2±0.9
Model 3	Yok	71.2±0.3	72.6±0.9	34.4±0.7	38.1±0.7
(k=4)	Dl=1	71.2±0.4	72.6±1.1	34.5±07	38.3±0.6
	Dl=1 Dr=1	**72.3**±0.3	**73.5**±1.0	35.5±0.9	38.7±0.9
	Hl=1 Hr=1	55.2±0.3	55.1±0.7	22.0±0.6	24.1±0.6
	Dl=1 Dr=1 Hl=1 Hr=1	71.1±0.3	72.4±0.9	35.5±0.8	38.4±0.9

Tablo 2. Daha küçük boyutlu eğitim verisi kullanmanın etkisi

		ÇK-ÇK bağlılıklarının Doğru olma yüzdesi		Tüm bağlılıkları doğru olan tümcelerin yüzdesi	
Model	Bağlam	Sözcük+Nokt.	Sözcük	Sözcük+Nokt.	Sözcük
Model 3	Yok	71.0±0.6	72.2±1.5	34.4±1.0	38.1±1.1
(k=4, 1500 Tümce)	Dl=1 Dr=1	71.6±0.4	72.6±1.1	35.1±1.3	38.4±1.5

Tablo 3. Sözcük-Sözcük doğruluğu değerlendirmesi sonuçları

		Sözcük-Sözcük Bağlılıkları doğru olma yüzdesi
Model	Bağlam	Sadece Sözcük
Dayanak 1	-	72.1±0.5
Dayanak 2	-	72.1±0.5
Dayanak 3	-	80.3±0.7
Model 1 (k=4)	Dl=1 Dr=1 Hl=1 Hr=1	80.8±0.9
Model 2 (k=5)	Dl=1 Dr=1	81.0±0.7
Model 3 (k=4)	Dl=1 Dr=1	**81.2**±1.0
Saf sözcük tabanlı	Yok	77.7±3.5

Tablo 4. Farklı tümce uzunlukları üzerinde başarım

Tümce uzunluğu l (ÇK'ler)	% Başarım
$1<l\leq10$	$80.2_{\pm0.5}$
$10<l\leq20$	$70.1_{\pm0.4}$
$20<l\leq30$	$64.6_{\pm1.0}$
$30<l$	$62.7_{\pm1.3}$

lığın modellenmesinde farklı yaklaşımlar veya kısmi görünüm bilgisi eklenmesinin yararlı olabileceğini belirtmektedir.

Sonuç

Bu makalede, Türkçe ağaç yapılı derlemdeki tümceler ile eğittiğimiz olasılık tabanlı modelleri kullanarak elde ettiğimiz Türkçe'nin olasılık tabanlı bağlılık ayrıştırması sonuçlarımız sunulmuştur. Bağlılık ilişkileri çekim kümeleri (ÇK) adını verdiğimiz sözcüklerden daha küçük birimler arasında kurulmakta ve ayrıştırıcımız bu ÇK'ler arasındaki bağlılıkları bulmaktadır. Derlemin küçük boyutu gözönüne alınarak, görünüm bilgisi içermeyen, ÇK'lerin biçimbirimsel özelliklerinin indirgenmiş gösterimlerini kullanan olasılık tabanlı modeller geliştirilmiştir. Bu çalışmanın amacına uygun olarak, incelemeler sağa bağımlı türde kesişmeyen bağlılıkları ayrıştırmakla sınırlandırılmıştır. En iyi sonuçlar (%73.5 ÇK-ÇK başarımı) ayrıştırma birimi olarak ÇK'lerin ve bağlam bilgisi olarak sözcük sonu ÇK'lerinin kullanıldığı bir model ile elde edilmiştir. Gelecek çalışmalar hataların daha ayrıntılı olarak incelenmesini, kısmi görünüm bilgisi kullanımının ve karar destek makineleri, bellek tabanlı öğrenme yöntemleri gibi daha karmaşık modellerin incelenmesini gerektirmektedir.

Teşekkür

Bu çalışma, TÜBİTAK Bilim İnsanı Destekleme Daire Başkanlığı ve İTÜ Bilimsel Araştırma Projeleri birimi tarafından desteklenmiş ve EACL`06'da yayınlanmıştır (Eryiğit ve Oflazer, 2006).

Kaynaklar

Charniak, E. (2000). A maximum-entropy inspired parser, *Proceedings,* 1st Conference of the North American Chapter of the Association for Computational Linguistics, 132-139, Seattle WA.

Chung, H. ve Rim, H. (2004). Unlexicalized dependency parser for variable word order languages based on local contextual pattern. *Proceedings,* Computational Linguistics and Intelligent Text Processing, 109-120, Seoul.

Collins, M., Hajic, J., Ramshaw, L. ve Tillmann, C. (1999). A statistical parser for Czech. *Proceedings*, 37th Annual Meeting of the Association for Computational Linguistics (ACL), 505–518, Maryland.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. *Proceedings*, 34th ACL, 184-191, Santa Cruz, CA.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. *Proceedings*, 35th ACL, 16–23, San Francisco.

Eryiğit, G. ve Oflazer, K. (2006). Statistical dependency parsing of Turkish. *Proceeding,* 11th Conference of the European Chapter of the Association for Computational Linguistics, 89-96, Trento.

Klein, D. ve Manning, C. (2003). Accurate unlexicalized parsing. *Proceedings*, 41st ACL, 423–430, Sapporo.

Kudo, T. ve Matsumoto, Y. (2000). Japanese dependency analysis based on support vector machines. *Proceedings*, Empirical Methods In Natural Language Processing and Very Large Corpora, 18-25, Hong Kong.

Kudo, T. ve Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. *Proceedings*, 6th Conference on Natural Language Learning, 63-69, Taipei.

Nivre, J. ve Nilsson, J.. (2005). Pseudoprojective dependency parsing. *Proceedings*, 43rd ACL, 99–106, Ann Arbor MI.

Nivre, J., Hall, J. ve Nilsson, J. (2004). Memory-based dependency parsing. *Proceedings*, 8th Conference on Computational Natural Language Learning, 49-56, Boston MA.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. *Proceedings*, 8th

International Workshop on Parsing Technologies, 23–25, Nancy.

Oflazer, K., Say, B., Hakkani-Tür, D. ve Tür, G. (2003). *Building a Turkish treebank* in Abeille, A. eds, *Building and Exploiting Syntactically-annotated Corpora*. Kluwer Academic Publishers, 261-277.

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

Oflazer, K. (2003). Dependency parsing with an extended finite-state approach. *Computational Linguistics*, **29,** 4, 515-544.

Sekine, S., Uchimoto, K. ve Isahara, H. (2000). Backward beam search algorithm for dependency analysis of Japanese. *Proceedings*, 17th International Conference on Computational Linguistics, 754–760, Saarbrücken.

Yamada, H. ve Matsumoto, Y. 2003. Statistical dependency analysis with support vector machines. *Proceedings*, 8th International Workshop of Parsing Technologies, 195-206, Nancy.