Türkçe Derlemler İçin Söz Dizimsel Görselleştirme ve Sorgulama Aracı

Syntactic Visualisation and Query Tool for Turkish Corpuses

Cem Agan, Banu Diri
Yıldız Teknik Üniversitesi Bilgisayar Müh. Bölümü, İstanbul
cemagan@hotmail.com, banu@ce.yildiz.edu.tr

Öz

Bu çalışmada, derlem, derlem türleri, mevcut Türkçe derlemler ve Türkçe bir derlemin etiketlenmesi gibi konular üzerinde durulmuştur. Ayrıca, Türkçe derlemlerden etkili bir şekilde faydalanmayı sağlayacak sorgulamalara imkan tanıyan ve Türkçe derlemlerdeki cümlelerin söz dizim ağaçlarını sözcük türleriyle birlikte görselleştiren bir araç geliştirilmiştir.

Abstract

In this study, we discuss subjects like text corpus, corpus types, available Turkish corpora and labeling a Turkish corpus. Besides, we offer a tool visualising the syntax trees of the sentences along with part of speech in a Turkish corpus and enabling queries which make it possible to efficiently make use of a Turkish corpus.

1. Giriş

Dil bilimde, büyük miktarda metin içeren yapılandırılmış doküman kümelerine derlem denir. Günümüzde çoğunlukla elektronik ortamda saklanır ve işlenirler. Derlemler dile ait istatistiksel çözümlemeler ve hipotez testleri (bir istatistiksel çıkarım metodudur) yapmada kullanılırlar.
Derlemler, dilbilimsel araştırmalara elverişli hale getirilmek için etiketleme (annotation) denilen bir işleme tabi tutulurlar. Bu işlemle derlemdeki metinsel bilgiye dilbilimsel bazı açıklamalar eklenir. Etiketlemeye örnek olarak sözcük türü etiketleme gösterilebilir. Sözcük türü etiketlemede her bir sözcüğün türü (fiil, isim, sıfat, vs.) etiketler şeklinde derleme eklenir. Başka bir örnek olarak da sözcük köklerinin belirtilmesi verilebilir.

Bazı derlemler tam kapsamlı dilbilimsel incelemeden geçirilmiş olabilir. Bunlar ağaç yapılı derlem olarak da adlandırılır ve söz dizimsel veya anlam bilimsel açıdan cümle yapıları etiketlenmiştir. Böyle derlemler için bütün bir derlemin tamamen ve tutarlı bir şekilde etiketlendirildiğinin kontrol edilmesi zor olduğundan genellikle küçük boyutlu olmaları tercih edilir.

Derlemler, derlem dilbilimindeki temel bilgi tabanlarıdır. Farklı türlerdeki derlemlerin incelenmesi ve işlenmesi hesaplamalı dilbilim, konuşma tanıma ve makine çevirisi alanlarındaki bir çok çalışmanın konusudur. Genellikle Saklı Markov modellerinin oluşturulmasında kullanılırlar. Bu modeller sözcük türü etiketleme ve diğer amaçlar için kullanılır. Bunlardan türetilen yeni derlem ve frekans listeleri dil öğretiminde de kullanılır. Derlem aynı zamanda yabancı dilde yazı yazmaya yardımcı bir unsur olarak da düşünülebilir [1].

Kısaca bu çalışmada, Türkçe bir derlem oluşturup mevcut dilbilimsel analiz araçlarıyla etiketlenmesi üzerine örnek bir uygulamaya yer verilmiş ve oluşturulan test derlemlerini söz dizimsel ilişkiler yönünden görselleştirip bu derlemler üzerinde karmaşık sorgular yapmayı sağlayan yeni bir aracın gerçekleştirme aşamaları tanıtılmıştır.

Makalenin ikinci bölümünde mevcut Türkçe derlemler hakkında genel bilgiler verilmiştir. Üçüncü bölümde test derleminin etiketlenmesi, dördüncü ve beşinci bölümlerde ilişki tablosu ve diyagramların çıkarılmasından, altıncı bölümde ise graf veri tabanı ve sorgu modülünün oluşturulmasından, yedinci bölümde de geliştirilen aracın kullanım alanlarından bahsedilmiştir.

Gönderim ve kabul tarihi : 08.12.2015 - 19.04.2016

2.Türkçe Derlem Çalışmaları

Türkçe üzerine yapılmış derlem çalışmalarının ilklerinden biri ODTÜ Türkçe Derlemidir [2]. 1990 sonrasına ait yazılı Türkçe örneklerden oluşmuştur ve 2 milyon sözcüğe sahiptir. Dokümanlar 10 değişik türden alınmıştır. Derlem XCES [3] etiketleriyle işaretlenmiştir. Araştırmacılar kullanıcı anlaşması formunu doldurmak şartıyla bu derlemden faydalanabilmektedir [4].

ODTÜ derleminin bir alt kümesi (7262 cümle), biçim birimsel ve söz dizimsel etiketlerle işaretlenerek ODTÜ-Sabancı Ağaç Yapılı Türkçe Derlemi [5] oluşturulmuştur. Bu derlemde yazı türlerinin dağılım oranı ODTÜ Türkçe Derlemine yakın tutulmuştur. Derlemin yapısı XML tabanlıdır. Araştırmacılar, aynı ODTÜ Türkçe derlemine benzer şekilde bu derlemden de faydalanabilmektedir.

Literatürde geçen bir diğer derlem çalışması Boğaziçi Üniversitesi akademisyenlerince hazırlanan BOUN Corpus'tur [6]. Bu derlem dört alt derlemden meydana gelmektedir. Bunların üçü Türkçe gazetelerden (NewsCor), biri de Türkçe web sayfalarından (GenCor) örneklenmiştir. Derlem toplamda 423 milyon sözcükten oluşmaktadır.

Türkçe Ulusal Derlemi (TUD) [7] 50 milyon sözcükten oluşan, 20 yıllık bir dönemi (1990-2009) kapsayan, günümüz Türkçesinin çok sayıda farklı alan ve türlerinden yazılı ve sözlü örneklerini içeren, geniş kapsamlı, dengeli ve temsil yeterliliğine sahip, genel amaçlı bir referans derlemdir. TUD-Tanıtım Sürümü, 1990-2009 yıllarını kapsayan 4438 veri kaynağından seçilen, 9 konu alanını ve 34 dilsel türü içeren doküman örneklerinden oluşmaktadır. Kullanıcılar yaklaşık 48 milyon sözcük üzerinden, medya, metin örneklemi, konu alanı, türev metin biçimi, yazar cinsiyeti, yazar türü, hedef okur ve metin türü
kısıtlama ölçütleriyle sorgularını gerçekleştirebilirler [8].
48 milyon sözcük üzerinden, medya, metin örneklemi, konu alanı, türev metin biçimi, yazar cinsiyeti, yazar türü, hedef okur ve metin türü kısıtlama ölçütleriyle sorgularını gerçekleştirebilirler [8].

İlk versiyonu 1 Mart 2012'de yayınlanan TS Corpus [9] ise, tamamı sözcük türü ve biçim birimsel bazda işaretlenmiş toplam 491 milyon birimden (491.360.398 adet token) oluşan genel amaçlı bir Türkçe derlemdir. Web sayfasından çevrimiçi erişime açıktır.

3.Test Derlemlerinin Etiketlenmesi

Çalışmaya ilk olarak küçük test derlemleri oluşturularak başlanmıştır. Derlemleri otomatik etiketleyeceğimiz ve etiketleme doğruluğunun yüksek olmasını istediğimiz için sözcük kökü çeşitliliğinin düşük ve cümle yapılarının göreceli olarak basit olduğu dokümanlardan örnekler kullanılmıştır. Buna göre Milli Eğitim Bakanlığı tarafından 2012 yılında yayınlanan Hayat Bilgisi 1 [10] ve Hayat Bilgisi 2 [11] ders kitapları OCR'den geçirilerek ham metinleri elde edilmiştir. Sonra her metin imla denetiminden geçirilmiş ve metinler cümlelere ayrılmıştır. Cümle segmentasyonunda basit regex ifadeleri yeterli olmuştur. Test derlemlerinin özellikleri Çizelge 1'de yer almaktadır.

Çizelge-1:Test Derlemlerinin Özellikleri

Derlem	Sözcük Sayısı	Cümle Sayısı
Hayat Bilgisi 1	3207	655
Hayat Bilgisi 2	6576	1246

Çalışmamız için sözcük türü ve söz dizim etiketleri gerekmekteydi. Bu ihtiyacımızı Turkish NLP Pipeline [12] araçlarını kullanarak giderdik.

Çalışmamızı yaptığımız sırada bu araç kümesinin 2013 sürümü elimizde mevcuttu. Kısa süre sonra bu araçlar http://tools.nlp.itu.edu.tr/ adresinden çevrimiçi kullanıma açılmıştır.

Turkish NLP Pipeline üç ayrı modülden oluşmaktadır. İki seviyeli biçim birimsel çözümleyici [13], algılayıcı tabanlı biçim birimsel bulanıklık giderici [14] ve bağımlılık ayrıştırıcı [15]. NLP Pipeline ile derlemimizi etiketlerken öncelikle doküman içerisindeki her birimin (sözcük veya noktalama) bir satırda ve cümle sonlarında da beş adet yıldız karakteri olacak şekilde bir giriş dosyasına yazılması sağlanmıştır. Bu dosya pipeline tarafından önce biçim birimsel çözümlemeye tabi tutulmuş, sözcüklerin kök ve ekleri belirlenmiştir. Bu işlemin çıktısı biçim birimsel bulanıklık gidericiye giriş olarak aktarılmıştır. Her sözcük için en olası biçim birimsel çözümleme seçilmiş ve bu işlemin çıktısı da söz dizimsel çözümleyiciye giriş olarak verilmiştir.

Şekil-1: Örnek bir cümle için etiketleme işlemi

Bu aşamada sözcüklerin hangi sözcükle söz dizimsel ilişki içerisinde olduğu referans numaraları ile belirtilmiştir.

Şekil-1'de 'Ömer, çantasına görsel sanatlar dersi için resim defteri ve boya kalemlerini koydu.' cümlesi için etiketleme işleminin basamakları gösterilmiştir. Şekil-1'deki son kutudaki analizler CoNLL [16] formatında verilmiştir. Sütunlarda sırasıyla birim numarası, sözcük-noktalama, sözcük kökü, sözcük türü, sözcük türü alt grubu, biçim birimsel özellikler, referans edilen sözcük numarası ve bağımlılık ilişkisi yer almaktadır.
Tüm cümleler etiketlendikten sonra aralarına ayırıcı bir kod konularak özel bir formatta derlem dosyası halinde kaydedilmiştir.

4. İlişki Tablosu

Örnek derlemlerimiz için, etiketli derlem dosyaları oluşturulduktan sonra bu etiketlerin doğruluklarının sınamasına geçilmiştir. Bu iş yapılırken her cümle için Şekil-1'de en altta yer alan kutudaki biçim birimsel ve söz dizimsel analizlerin doğrulanması gerçekleştirilmiştir. Biçim birimsel analizler bu kutudan kolayca takip edilebilmesine karşın, söz dizimsel analizlerin takibi cümle yapısı karmaşıklaştıkça zorlaşmaktadır. Bu takibi kolaylaştırmak için aracımıza bir 'ilişki tablosu' modülü eklenmiştir. Bu sayede derlemden her hangi bir cümle seçildiğinde bilgisayar tarafından sözcükler arasındaki söz dizimsel ilişkileri kuş bakışı gösteren iki boyutlu bir grafik üretilmektedir. Bu grafikte sütunlarda sözcükler, satırlarda da (ilişki katmanlarında) ilişki tipleri yer almaktadır. İlişkiler ise baklava parçaları ile gösterilmektedir. Baklavaları oluşturan ilk üçgene (hücrede sağa yapışık) ait sözcük referans veren, ikinci üçgene (hücrede sola yapışık) ait sözcük de referans alan sözcüktür.

Şekil 2'de bu grafiğe bir örnek verilmiştir. Satırlardaki ilişki tipi isimleri kısaltma şeklindedir ve Çizelge 2'de kısaltmaların açıklamaları verilmiştir.

Çizelge-2:İlişki Tablosundaki Kısaltmaların Açıklamaları

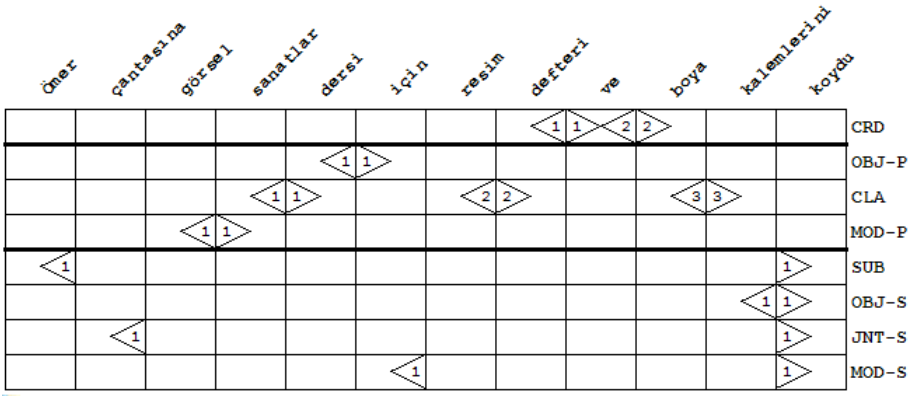
Kısaltma	Açıklama
CRD	Coordination
CLA	Classifier
SUB	Subject
OBJ	Object
JNT	Adjunct
MOD	Modifier

Ayrıca, bazı ilişki tipleri için hem TİP-P hem de TİP-S versiyonları mevcuttur. İlişkinin bir öbek (P) mi yoksa bir cümle (S) mi oluşturduğuna bakılarak hangi versiyonun kullanılacağı belirlenir.

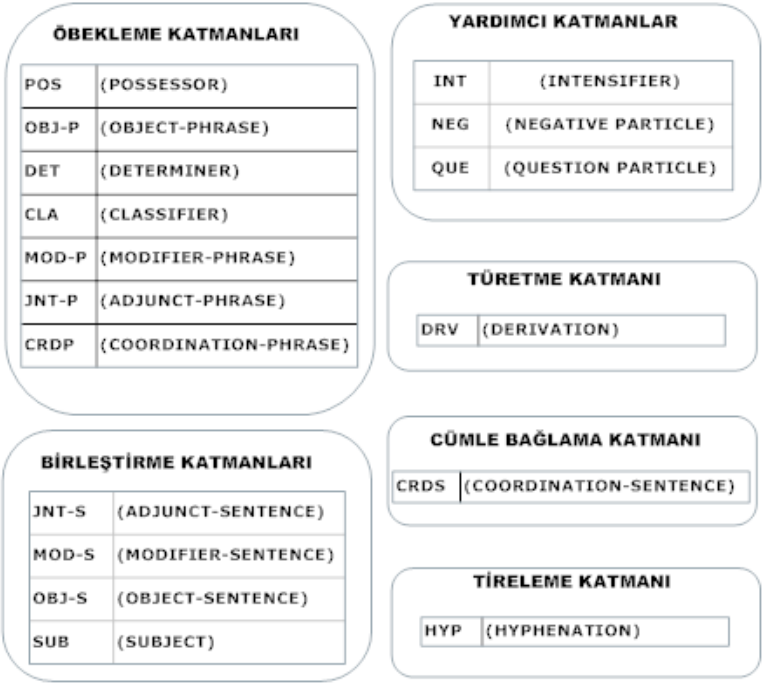
Bir ilişki tablosu için olası ilişki katmanı grupları ve bunlara ait ilişki katmanları Şekil 3'deki gibi belirlenmiştir. Öbekleme katmanındaki ilişkiler sözcük öbeklerini oluşturmaktadır. Birleştirme katmanındaki ilişkiler ise cümleyi meydana getiren en büyük yapıları birleştirerek cümleyi oluşturmaktadır. Türetme katmanı sözcüklerdeki türeme ilişkilerini göstermekte, cümle bağlama katmanı ise cümleleri bağlaçlarla birbirine bağlamaktadır.

Örnek derlemimizdeki cümlelerin analizleri, otomatik oluşturulan ilişki tabloları yardımıyla kontrol edilmiştir. Basit yapılarından dolayı örnek derlemlerimizdeki cümlelerin söz dizimsel analizlerinin hata oranları düşük çıkmıştır.
Söz dizimsel etiketlemede karşılaştığımız hataların en önemli nedeni yanlış biçim birimsel çözümlemenin seçilmiş olmasıdır. Örneğin 'yazınız', 'veriniz' ve 'düşünürler' gibi fiillerin isim olarak algılanması söz dizimsel analizin yanlış yapılmasına neden olmuştur:

yazınız = yazı+Noun+A3sg+P2pl+Nom
veriniz = veri+Noun+A3sg+P2pl+Nom
düşünürler = düşünür +Noun +A3pl+Pnon+Nom



Şekil-2: Örnek bir cümle için ilişki tablosu



Şekil-3: Bir ilişki tablosunda yer alabilecek ilişki katmanları5. İlişki Diyagramı

İlişki tablolarından sonra sözcükten öbeklere, cümleciklere ve cümlelere kadar her söz dizimsel düzeyden elde edilen yapıları gösteren bir ilişki diyagramı modülü oluşturulmuştur. İlişki diyagramları oluşturulurken ilişki tablosundan faydalanılmıştır. Şekil 4'de bir önceki örnek cümlemiz için üretilen ilişki diyagramı görülmektedir.

Diyagramda dört isim öbeği, bir edat öbeği (tümleç durumunda), bir bağlaç öbeği (nesne durumunda) vardır. Kısaltmaların anlamları Çizelge 3'de verilmiştir.

Çizelge-3: Diyagramdaki Kısaltmalar

Kısaltma	Açıklama
O	Özne
N	Nesne
Y	Yüklem
T	Tümleç
I	İsim
S	Sıfat
B	Bağlaç
F	Fiil
C	Cümle
IK	İsim Öbeği
EK	Edat Öbeği
BK	Bağlaç Öbeği
NT	Niteleyici
BS	Baş sözcük
BG	Bağlanan
BY	Bağlayıcı

Şekil 4'deki gibi bir graf yapısı elde edilirken aşağıdaki işlemler gerçekleştirilir:

1. Her sözcük için öbekleme katmanlarındaki tüm katmanlar yukarıdan aşağıya doğru taranır, eğer katmanda sözcüğün referans verdiği bir ilişki varsa bu ilişkiyi oluşturan sözcükler öbekleme kurallarına göre öbeklenir.
2. Birleştirme katmanlarındaki tüm ilişkiler işlenerek birleştirme kurallarına göre cümle oluşturulur.
3. Türetme katmanındaki tüm ilişkiler işlenir.
4. Yardımcı katmanlardaki tüm ilişkiler işlenerek yardımcı katman kurallarına göre gerekli öbekler oluşturulur.
5. Cümle bağlama katmanındaki tüm ilişkiler işlenerek cümleler bağlaçlarla birbirine bağlanır.

Sözü geçen kurallar graf oluşturma kurallarıdır. Bir cümle için oluşturmakta olduğumuz grafa, *a* ve *b* düğümlerini (sözcükleri) ekleyeceğimizi ve bunların a→b şeklinde bir söz dizimsel ilişkiye sahip olduğunu varsayalım. Buna göre kurallar Çizelge 4'deki gibidir. Burada n, n_1, n_2 herhangi bir düğümü (her düğüm bir söz dizimsel birimi temsil eder), *n*' yeni oluşan bir düğümü, kırmızı (koyu renk) ok iptal edilen bir ilişkiyi, yeşil ok (açık renk) yeni kurulan bir ilişkiyi, EBA(x) ise *x* düğümünün en büyük atası olan düğümü ifade etmektedir.

Şekil-4: Örnek bir cümle için ilişki diyagramı

Graf oluşturma kurallarının başarımı test derlemleri üzerinde cümle bazında ölçülmüştür. Çizelge 5'de diyagramı doğru çıkarılan cümlelerin sayıları ve oranları verilmiştir.

Çizelge-5:Graf Oluşturma Kurallarının Başarım

Derlem	Doğru Diyagram Sayısı	Başarı Oranı
Hayat Bilgisi 1	594	%90.68
Hayat Bilgisi 2	1165	%93.49

Tablodaki sayılar söz dizimsel diyagramı tamamıyla doğru çıkarılmış cümle sayılarını göstermektedir. Bir cümlenin diyagramında kısmi bir hata dahi olsa bu diyagram hatalı kabul edilmiştir.

Çizelge-4: Söz dizimsel ağaç oluşturma kuralları

ÖBEKLEME KURALLARI	
Durum	İşlem
a→b bağlantısı yapılacak. a ve b grafta yok.	a→n', b→n'
a→b bağlantısı yapılacak. grafta a yok, b →n var.	b→n, a→n', b→n', n'→n
a→b bağlantısı yapılacak. grafta a→n var, b yok.	EBA(a)→n', b→n'
a→b koordinasyon bağlantısı yapılacak. grafta $a \rightarrow n_{1(crd\ öbeği)}$ ve $b \rightarrow n_2$ var.	$b \rightarrow n_2$, $b \rightarrow n_1$, $n_1 \rightarrow n_2$

ÖBEK BAĞLAMA KURALLARI	
Durum	İşlem
a→b bağlantısı yapılacak. a ve b grafta yok.	a→n', b→n'
a→b bağlantısı yapılacak. grafta a→n var, b yok. a'nın tipi noktalama veya bağlaç değilse: aksi durumda:	EBA(a)→n', b→n' b→n
a→b bağlantısı yapılacak. grafta $a \rightarrow n_1$ ve $b \rightarrow n_2$ var.	EBA(b)→n_1

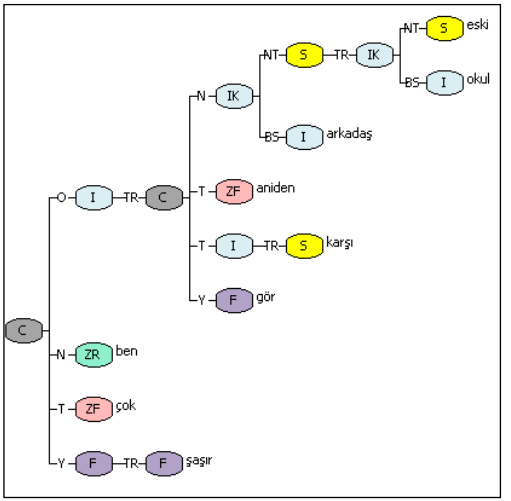
BİRLEŞTİRME KURALLARI	
Durum	İşlem
a→b bağlantısı yapılacak. a ve b grafta yok.	a→n', b→n'
a→b bağlantısı yapılacak. grafta a→n var, b yok.	EBA(b)→n
a→b bağlantısı yapılacak. grafta a yok, b→n var.	a→n
a→b bağlantısı yapılacak. grafta $a \rightarrow n_1$ ve $b \rightarrow n_2$ var.	EBA(a)→n_2

TÜRETME KURALLARI	
Durum	İşlem
a→b bağlantısı yapılacak. grafta a yok, b→n var.	a→b
a→b bağlantısı yapılacak. grafta $a \rightarrow n_1$ ve $b \rightarrow n_2$ var.	n_1→b

CÜMLE BAĞLAMA KURALLARI	
Durum	İşlem
a→b bağlantısı yapılacak. grafta $a \rightarrow n_1$ ve $b \rightarrow n_2$ var. eğer a bağlaç ise ise; değilse;	EBA(b)→ EBA(a) EBA(a)→ EBA(b)
a→b bağlantısı yapılacak. grafta a→n var, b yok.	EBA(a)→n', b→n'

Bağımlılık çözümlemesinde ilişkiler bazen sözcük türemelerinin olduğu yerlerde de bulunmaktadır. Bu tür durumlarda diyagramlar üretilirken önce birbirleriyle ilişkili gruplar diyagram üzerinde birbirlerine bağlanır sonra da 'TR' kısaltması eklenerek sözcük türetimi gerçekleştirilir ve diyagrama kalındığı yerden devam edilir.

Şekil 5'teki diyagramda bu duruma bir örnek verilmiştir. Diyagram 'Eski okulumdaki arkadaşlarımı aniden karşımda görmek beni çok şaşırttı.' cümlesine aittir. Ağacın kökündeki C düğümü ana cümleciği gösterirken, diğer C düğümü yan cümleciği temsil etmektedir. TR kısaltmaları ise türemeleri göstermektedir. Bu cümlede 'eski okulumda' grubuna –ki eki getirilerek bir sıfat oluşmuştur. Ayrıca, yan cümlecik bir isim oluşturmuş ve ana cümleciğin öznesini meydana getirmiştir.

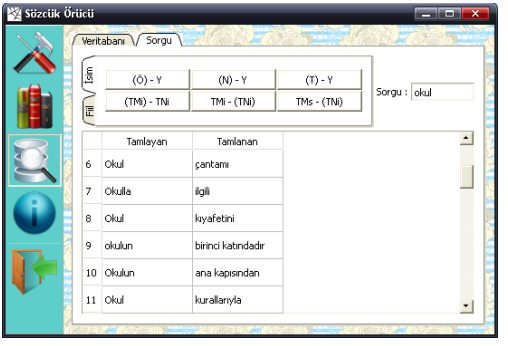


Şekil-5: Türemelerin olduğu yerlerde ilişkilerin gösterilmesi

6. Graf Veri tabanı ve Sorgu Modülü

Derlemler üzerinde söz dizimsel sorgular yapılabilmesi için derlemlerdeki cümlelerin ilişki grafları bir graf veri tabanına kaydedilmiştir. Bunun için yüksek oranda bağlantılı verileri saklamakta kullanılan bir graf veri tabanı olan Neo4j [17] kullanılmıştır. Neo4j, verileri düğümler ve bağlantılar şeklinde saklayan açık kaynaklı bir graf veri tabanı olup, düğümlere ve bağlantılara etiket ve özellikler verilebilmektedir.

Graf veri tabanından sorgulama yapabilmek için bir sorgulama modülü (Şekil-6) yazılmıştır. Bu modülle isim ve fiil kökleri anahtar sözcük olmak üzere aşağıdaki aramalar yapılabilmektedir:



Şekil-6: Örnek bir cümle için etiketleme işlemi

- anahtar sözcüğün (isim) nesne olduğu cümlelerin fiilleri
- anahtar sözcüğün (isim) tümleç olduğu cümlelerin fiilleri
- anahtar sözcüğün (isim) tamlayan olduğu isim tamlamalarının tamlananları
- anahtar sözcüğün (isim) tamlanan olduğu isim tamlamaların tamlayanları
- anahtar sözcüğün (isim) tamlanan olduğu sıfat tamlamalarının tamlananları
- anahtar sözcüğün (fiil) yüklem olduğu cümlelerin özneleri
- anahtar sözcüğün (fiil) yüklem olduğu cümlelerin nesneleri
- anahtar sözcüğün (fiil) yüklem olduğu cümlelerin tümleçleri

Şekil-6'daki sorgulama arayüzünden öntanımlı sorgular yapılabilmektedir. Ancak, graf veritabanı servisinin web arayüzündeki sorgulama ekranından Neo4j'in sorgulama dili olan Cypher [18] ile oldukça detaylı sorgular da yapmak mümkündür. Örneğin, 'Öğretmenlerin yaptığı eylemleri (*olmak* ve *var* hariç) yapan diğer özneleri bul' sorgusu Cypher dilinde '*MATCH (a)-[:O]-(cumle)-[:Y]-(b), (c)-[:Y]-(cumle2)-[:O]-(d) WHERE a.kok="öğretmen" and NOT b.kok IN ["_", "ol", "var"] and c.kok=b.kok return b,c,d.kok;*' şeklinde yazılabilir. Bu sorguya cevap olarak Çizelge-6'daki gibi kayıtlar dönmektedir.

Çizelge-6: Karmaşık sorgunun dönüş değerler

b	c	d.kok
çalışır	çalıştı	Atatürk
çalışır	çalışır	Hakim
kullanmalıdır	kullanırlar	Doktor
...	...	...

Bu sorgu 'Şu beş koşulu sağlayan, b ve c sözcüklerini ve d sözcüğünün kökünü bul. 1. a özne olarak, b de yüklem olarak bir cümleye bağlı 2. c özne olarak, d de yüklem olarak bir cümleye bağlı 3. a'nın kökü öğretmen 4. b'nin kökü {olmak,var} kümesinde yok 5. c'nin kökü b'nin kökü ile aynı.' şeklinde de okunabilir.

7. Kullanım Alanları

Ne kadar çok ham veri detaylı dilbilimsel çözümlemeye tabi tutulup sözcükler ilişkilendirilirse bu ilişkilerin frekansları gerçek hayattaki doğal değerlerine yaklaşır. Bu ilişkiler daha gerçekçi Doğal Dil İşleme uygulamalarının geliştirilmesine imkan verecektir.

Sözcükler arası ilişkilerle örülmüş büyük bir veri yığının pratikte birçok faydası vardır. Bunlara ilk olarak bağlama duyarlı uygulamalar örnek verilebilir. Bir paragraf veya yazının içerisindeki cümlelerde geçen sözcüklerin bağlam modelleme için oluşturulmuş bir graf üzerinde birleştiği yollara yüksek frekanslarla bağlı olan sözcükler, semantik yorumlama sırasında arama alanını daraltmakta kullanılabilir.

Bir önceki bölümdeki gibi graf veri tabanına karmaşık sorgular yaparak birbirine benzer kavramları gruplamak ontoloji çıkartmaya yardımcı olabilir. Cümlelerde geçen varlıklara ait özellikler ve bu varlıkların sergiledikleri eylemler göz önüne alınarak varlıklar arasındaki benzerlikler ölçüldüğünde kümeleme yöntemiyle, aynı hiyerarşik seviyede olan kavramlar tespit edilebilir. Ayrıca, sözcük ağı oluşturma projelerinde belirli semantik ilişkilerin yakalanmasında graf veri tabanlarından faydalanılabilir. Bazı işlevsel (dil bilgisel) sözcüklerin oluşturdukları öbekler sorgulanarak sözcükler arasındaki semantik ilişkiler ortaya çıkartılabilir.

Başka bir uygulama alanı olarak da akıllı sözcük önerileri veren ve yazımdaki bazı mantıksal hataları düzelten uygulamalar verilebilir. Örneğin, metin girişi içeren bir uygulamada sözcükler girilirken dinamik olarak bu sözcüklerle yüksek frekansta söz dizimsel ilişkiye sahip olan sözcükler uygulama tarafından kullanıcıya önerilerek hem hızlı hem de yerinde kelime seçimleri mümkün kılınabilir. Ayrıca, metinlerdeki dil bilgisel yönden doğru fakat anlamsal yönden gerçekten uzak olan hatalı sözcükler tespit edilip yerine doğru sözcük önerileri verilebilir. Örneğin, 'Güneşi *atmayan* imparatorluk' ifadesinde geçen atmak fiili imla yönünden doğru olmakla birlikte anlamsal yönden gerçeğe uygun değildir. Güneş varlığının özne olarak geçtiği cümle veya cümleciklerin fiilinin 'atmak' olma olasılığı oldukça düşüktür. Söz dizimsel ilişkileri barındıran derlemlerden üretilen graf veri tabanlarına yapılacak sorgulamalarda bu durum tespit edilip, olası eylemlerden bu fiile biçimsel olarak en yakın ve frekans olarak en yüksek olan 'batmak' kelimesi yazım önerisi olarak verilebilir.

8. Sonuç

Derlemler yapılandırılmış büyük metin kümeleridir ve dilbilimsel araştırmalarda kullanılabilmeleri için sözcük türü, biçimbirim, sözdizim ve anlambilim etiketleriyle etiketlendirilirler. Hesaplamalı dilbilim ve doğal dil işleme gibi çalışma alanlarında yapılan araştırmalarda derlemlerden faydalanılır.
Türkçe derlem oluşturma çalışmaları 1990'lardan başlayıp günümüze kadar uzanmaktadır. ODTÜ Türkçe Derlemi, ODTÜ-Sabancı Ağaç Yapılı Türkçe Derlemi, BOUN Corpus, Türkçe Ulusal Derlemi ve TS Corpus mevcut Türkçe derlemlerin en bilinenleridir.

Makalede Türkçe derlemlerin söz dizimsel olarak görselleştirilmesi ve karmaşık sorgular yapılması üzerine gerçekleştirilen bir çalışmadan bahsedilmektedir. Bu çalışmada test amaçlı derlemler oluşturulmuş ve bu derlemler güncel dilbilimsel çözümleme araçlarıyla etiketlenmiştir. Çalışma kapsamında gerçekleştirilen bir araçla derlemdeki söz dizimsel ilişkiler otomatik olarak ilişki tabloları şeklinde görselleştirilmiştir. Ayrıca, bu tablolar yorumlanarak otomatik olarak cümlelerin söz dizim ağaçlarını gösteren ilişki diyagramları üretilmiştir. Araç yardımıyla derlemdeki söz dizimsel yapılar graflar halinde bir graf veri tabanına aktarılmış ve bir sorgulama arayüzü üzerinde öntanımlı bazı ilişkilerin sorgulanması sağlanmıştır. Çalışmada kullanılan graf veri tabanının web arayüzü üzerinden nasıl karmaşık sorgular yapılabileceğine dair bir örnek de verilmiştir.

Detaylı dilbilimsel etiketlere sahip derlemlerden faydalanılarak üretilen graf veri tabanları daha gerçekçi Doğal Dil İşleme uygulamalarının geliştirilmesine olanak sağlayacaktır. Bunlara örnek olarak bağlama duyarlı uygulamalardaki semantik yorumlama, ontoloji çıkartma ve semantik ilişkilerin keşfine yardımcı olan ve akıllı sözcük önerileri veren uygulamalar verilebilir.

Kaynakça

[1] Text Corpus, http://en.wikipedia.org/wiki/Text_corpus, 30.11.2014

[2] Say, B., Zeyrek, D., Oflazer, K., Özge, U., 2002. *Development of a corpus and a treebank for present-day written Turkish*, Proceedings of The Eleventh International Conference of Turkish Linguistics, s.183-192.

[3] Corpus Encoding Standard for XML, http://www.xces.org/, 30.11.2014

[4] Metu Turkish Corpus http://ii.metu.edu.tr/corpus, 30.11.2014

[5] Oflazer, K., Say, B., Hakkani-Tür, D.K., Tür, G., 2003. *Building a Turkish Treebank*, Invited chapter in "Building and Exploiting Syntactically-annotated Corpora", Anne Abeille Editor, Kluwer Academic Publishers.

[6] Sak, H., Güngör, T., Saraçlar, M., 2011. *Resources for Turkish morphological processing, Language Resources and Evaluation, 45(2), s.249-261.*

[7] Aksan, Y., 2012. *Construction of the Turkish National Corpus (TNC)*, In Proceedings of the Eight International Conference on Language Resources and Evaluation, İstanbul, Turkiye, s. 3223-3227.

[8] Türkçe Ulusal Derlemi (TUD), http://www.tnc.org.tr, 30.11.2014

[9] Sezer, B., Sezer, T., 2013. *TS Corpus: Herkes için Türkçe Derlem, Proceedings of the 27th National Linguistics Conference, Antalya, Hacettepe University, Linguistics Department, s. 217-225.*

[10] Dalkılıç H., Gölge N., 2012. MEB *İlköğretim Hayat Bilgisi Ders Kitabı 1*

[11] Özemir A., Çınar F., 2012. *MEB İlköğretim Hayat Bilgisi Ders Kitabı 2*

[12] Turkish NLP Pipeline, http://web.itu.edu.tr/gulsenc/pipeline.html, 30.11.2014

[13] Güngördü, Z., Oflazer, K., 1994. *Parsing Turkish using the Lexical-Functional Grammar Formalism*, In Proceedings of COLING'94, The 15th Conference on Computational Linguistics, Kyoto, Japan, s. 494–500.

[14] Sak, H., Güngör, T., Saraçlar, M., 2007. *Morphological Disambiguation of Turkish Text with Perceptron Algorithm*, In CICLing, LNCS 4394, s.107-118.

[15] Eryiğit,G., Nivre, J., Oflazer, K., 2008. *Dependency Parsing of Turkish*, Computational Linguistics, 34 (3), 357-389.

[16] CoNLL-X Shared Task: Multi-lingual Dependency Parsing, http://ilk.uvt.nl/conll/, 30.11.2014

[17] Neo4j, http://neo4j.org/, 30.11.2014

[18] Cypher Query Language, http://neo4j.com/developer/cypher-query-language/, 20.11.2014