TS Corpus: Herkes için Türkçe derlem

Taner Sezer[*]
Bengü Sever Sezer[**]
Mersin Ünivesitesi

Özet

TS Corpus'un ilk versiyonu 1 Mart 2012, ikinci versiyonu 30 Ağustos 2012'de yayınlanmıştır. TS Corpus 491M+ birimden oluşan, tamamı sözcük türü (PosTAG), biçimbirimsel yapı etiketi (Morphological Tagging) ve kök sözcük (Lemma) bazında işaretlemiş, CWB/CQP altyapısıyla oluşturulmuş bir Tükçe derlemdir. Bu nitelikleriyle BNC ve COCA gibi büyük veriye sahip ve modern derlemlerle ortak özellikleri sunmaktadır. İnternet üzerinden herhangi bir ücret veya onay talep edilmeden ulaşılabilen açık erişimli bir derlemdir. Bu çalışmada TS Corpus verisinin işaretleme öncesi hazırlanması ve işaretlenmesi sürecinden başlayarak, derlemin kullanıcıya ulaşmasına kadar olan süreç anlatılmaktadır. Geliştirme süreci devam eden TS Corpus'un gelecek sürümler için hedefleri ve TS Corpus'un Türkçe bilişimsel dilbilim çalışmalarına katkısı da ele alınacaktır. Bir sonraki sürümde daha fazla veri toplamak için veri toplayıcı (crawler) yazılımları ve üstmetin bilgilerine göre işlemek için makine öğrenme (machine learning) yöntemleriyle türlere ayrıştırma yoluyla 500 milyon–1 milyar arası sözcükten oluşan bir derlem oluşturmak hedeflenmektedir.

Anahtar Sözcükler: derlem, derlem dilbilim, corpus, sözcük türü işaretleme, doğal dil işleme

* Mersin Üniversitesi İletişim Fakültesi tanersezerr@gmail.com
** Mersin Üniversitesi Yabancı Diller Yüksekokulu bengusever@gmail.com

1. Çalışmanın Amacı

Derlemdilbilim ülkemizde son dönemde oldukça ilgi görmeye başlamış bir alandır. Bu alanda çeşitli çalışmalar yapılmakta, yeni derlemler oluşturulmakta ve kullanıcılara açılmaktadır. TS Corpus projesi, Türkçe için, öncülü ve çağdaşı diğer derlemlerde bulunmayan sözcük türü ve biçimbirimsel çözümleme etiketlerini içeren ve büyük bir veriye sahip bir derlem oluşturma amacıyla tasarlanmıştır. Hedeflenen ilk amaç British National Corpus XML Edition (BNCWeb XML) ile aynı arama ve istatistik özelliklerini Türkçe bir derlemde sunmaktır. Ancak çalışmanın asıl hedefi, bir derlem oluşturmanın ötesinde, farklı derlemler oluşturmak, bu amaçla ilgili alanda kullanılabilecek yeni betikler hazırlamak, karşılaşılan bilişimsel sorunları tanımlamaktır.

Çalışmada, Türkçe derlem dilbilim ve bilişimsel dilbilim alanında yapılmış çalışmalar ve bu çalışmalar sonucu elde edilen yazılım ve betikler ile dünyada derlem dilbilim alanında kullanılan yazılımlar bir araya getirilmeye çalışılmış, böylelikle birbirinden bağımsız gözüken ve dağınık halde duran bu çalışmalardan, kullanılabilir bir son kullanıcı ürünü ortaya çıkarmak hedeflenmiştir.

Bu çalışmanın yapıldığı dönemde henüz Türkçe için etiketlenmiş bir derlem olmayışı, sözcük türü ve biçimbirimsel olarak etiketlenmiş bir Türkçe derlemin eldeki mevcut çalışmalar ışığında yapılıp yapılamayacağı sorusunu akla getirmiştir. Bu bağlamda Türkçe verinin farklı katmanlarda işaretlenmesinin, bu işaretlenmiş verinin derlem arayüzünde kullanılacak yapıya taşınmasının yolları araştırılmıştır.

TS Corpus'un mevcut sürümleri dili temsil (representativeness) ve dengeli bir derlem (balance) olma iddaasında değildir. Üstmetin bilgilerine (metadata) göre arama yapma veya tür ve alanlara (genre&domain) göre sonuçların dağılımı görmek de TS Corpus'un mevcut sürümlerinde bulunan özellikler değildir.

Çalışmanın mevcut (aktif ve beta) sürümlerindeki hedefi sözcük türü (PosTag), biçimbirimsel çözümleme (morphological tagging) ve kök (lemma) yapılarını işaretleyerek, bu üç katmanda da arama yapılmasını sağlamaktır. TS Corpus sürüm1 ve 2 ile bu hedeflere ulaşılmış, Türkçe için ilk defa internet üzerinden erişilebilen, sözcük türü, biçimbirimsel etiket ve kök sözcük ile arama yapılabilen bir derlem erişime açılmıştır.

Ayrıca çalışmada kullanılan yazılımların tamamı açık kaynak kodlu yazılımlardır. Bu şekilde çalışmanın sürerliliği ve özgürlüğü hedeflenmiştir.

2. TS Corpus Nedir?

TS Corpus ilk sürümü 1 Mart 2012, ikinci sürümü 30 Ağustos 2012 tarihinde yayınlanmış olan, toplam 491M+ birimden oluşan, tamamı sözcük türü ve biçimbirimsel yapıda etiketlenmiş, CWB/CQPweb altyapısıyla oluşturulmuş, internet üzerinden erişilebilen, genel amaçlı bir Türkçe Derlemdir.[1] Çalışma evrenini iki başlık altında ele almak mümkündür. İlk parça derleme konu olan veri ve ikinci parça bu verinin işlenmesi ve sunulması amacıyla kullanılan yazılımlardır.

3. Veri

Çalışmada BOUN Web Corpus[2] verisi kullanılmıştır. XML biçimli iki ayrı dosya olarak internet üzerinden dağıtılan BOUN Web Corpus[3] toplam 4.4 Gb büyüklüğünde iki ayrı dosyadan oluşmaktadır. newscor.xml olarak anılan ve 1.9 Gb büyüklüğündeki ilk parça haber sitelerinden toplanmış verileri, 2.5 Gb büyüklüğündeki ikinci parça ise çeşitli diğer internet sitelerinden toplanmış veriyi içermektedir. Veri 491M+ birimden oluşmaktadır ve CES/XML olarak işaretlenmiştir.

---

1 TS Corpus sürüm 2 ile beraber TS Corpus Wikipedia -Beta- Derlemi de 30 Ağustos 2013 tarihinde erişime açılmıştır. Çalışma Türkçe Wikipedia sayfalarınından oluşturulmuş sözcük türü ve biçimbirimsel olarak işaretlenmiş 46M+ birimden oluşan bir veriyi içermektedir. http://tscorpus.com

2 Sak, H., Tunga, G., Saraçlar, M.. (2008) *Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus.* GoTAL 2008: p417-427.

3 http://79.123.177.209/~hasim/langres/BounWebCorpus.tgz

Corpus	Words	Tokens	Types	Tokens Parsed(%)
NewsCore	184M	212M	2.2M	96.7
GenCor	239M	279M	3.0M	94.6
BOUN Corpus	423M	491M	4.1M	95.5

Tablo 1. Sak, 2012:58

Tablo 1 BOUN Web Corpus'u oluşturan verinin dağılımını göstermektedir.

Verinin internet üzerinden dağıtılan biçimi her satıra bir sözcük (word-per-line) gelecek şekildedir. Her satır yalnızca ham olarak sözcüğün kendisini içermektedir. Biçimbirimsel çözümleme ve sözcük türü etiketleri bu veride bulunmamaktadır.

4. Yazılım

Çalışmada kullanılan betik ve yazılımlar ise açık kaynak kodlu özgür yazılımlar arasından seçilmiş, gerektiği noktalarda ihtiyacı karşılamak üzere yeni betikler hazırlanmıştır.

Sözcük türü işaretleme ve biçimbirimsel çözümleyici olarak olarak "An averaged perceptron-based morphological disambiguator for Turkish text"[4] kullanılmıştır.[5]

Ancak sözkonusu yazılımın çıktısı derlem altyapısını oluşturacak CWB/CQPweb[6] altyapısına uyumlu değildir. Uyumlu hale getirmek amacıyla yeni bir betik oluşturulmuştur. (TS Corpus CWB/CQPweb altyapısı ile hazırlanan ilk Türkçe derlemdir[7].)

Yazılımın Orijinal çıktısı ve hazırlanan betik ile alınan çıktısı tablo 2'de incelenebilir.

An averaged perceptron-based morphological disambiguator for Turkish text çıktısı	ts-parser betiği ile alınan değiştirilmiş çıktı
Geçen geç[Verb]+[Pos]-YAn[Adj+PresPart] : 8.6005859375 geçen[Adj] : 7.2607421875 geçe[Noun]+[A3sg]+Hn[P2sg]+[Nom] : 22.0634765625 hafta hafta[Noun]+[A3sg]+[Pnon]+[Nom] : 7.9365234375 haf[Noun]+[A3sg]+[Pnon]+DA[Loc] : 17.6328125 kısa kıs[Verb]+[Pos]+YA[Opt]+[A3sg] : 16.8857421875 kısa[Noun]+[A3sg]+[Pnon]+[Nom] : 12.0693359375 kısa[Adv] : 16.125 kısa[Adj] : 7.7890625 vadeli vade[Noun]+[A3sg]+[Pnon]+[Nom]-lH[Adj+With] : 11.7421875 vadeli[Adj] : 12.326171875 avans avans[Noun]+[A3sg]+[Pnon]+[Nom] : 12.4443359375 hesabına hesap[Noun]+[A3sg]+SH[P3sg]+NA[Dat] : 11.9873046875 hesap[Noun]+[A3sg]+Hn[P2sg]+NA[Dat] : 16.66015625 hesabına[Adv] : 15.431640625	Geçen Adj Adj geçen hafta Noun Noun+A3sg+Pnon+Nom hafta kısa Adj Adj kısa vadeli Noun Noun+A3sg+Pnon+Nom+Adj+With vade avans Noun Noun+A3sg+Pnon+Nom avans hesabına Noun Noun+A3sg+P3sg+Dat hesap

Tablo 2. İşaretlenmiş veri

Tablo 2'de sol sütun kullanılan yazılımın verdiği çıktıyı, sağ sütun hazırlanan betik ile elde edilen CWB/CQPweb uyumlu çıktıyı göstermektedir. Sağ sütunda görülen veri her satıra bir sözcük gelecek şekilde düzenlenmiş olup her bir nitelik (attribute) tab (düzenli ifade karşılığı \t) ile birbirinden ayrılmıştır.

Sol sütunda her bir girdi sözcük için olası bütün çözümlemeler verilmişken, hazırlanan betik tüm çözümlemeler içinden en olası etiketi seçmek üzere yazılmıştır. Bu yöntem sözcüğün bağlam içindeki sözcük türünü her zaman doğru seçmesi anlamına gelmemektedir. Başka bir deyişle sözcük türü ve biçimbirimsel çözümleme için anlam

4 *Sak, H., Tunga, G., Saraçlar, M.. (2007), Morphological disambiguation of Turkish text with perceptron algorithm.*. CICLing, Vol LNCS 4394, pp 107-118.

5 http://79.123.177.209/~hasim/langres/MD-2.0.tgz

6 http://cwb.sourceforge.net/

7 http://cwb.sourceforge.net/demos.php

bulanıklığı (disambiguation) giderilmemiştir. Ancak halihazırda erişilebilir bulunan sözcük türü ve biçimbirimsel çözümleyici yazılımları arasında en kullanabilir uygulama olarak çalışmada bu yol izlenmiştir. TS Corpus'un bu yöntemle sözcük türü etiketleme başarımı %80-82 aralığında görülmektedir.

Öte yandan girdi veri olarak kullanılan BOUN Web Corpus'un internet ortamından toplanan verilerden oluştuğu göz önünde bulundurulmalıdır. Dolayısıyla veri, ş,ğ, vb. karakterler yerine internet kullanıcıları tarafından tercih edilen s,g, vb. karakterler kullanılarak yazılmış sözcükler de içermektedir. Neşeli sözcüğü yerine (neseli, bağcı sözcüğü yerine bagci vb.) Veri, işaretleme öncesinde hiçbir yazım denetiminden geçerilmemiş, olduğu şekilde kullanılmıştır.

Derlemi oluşturan diğer temel bileşenler ise CWB/CQPweb[8], Apache web sunucusu[9], PHP5[10], MySQL sunucusudur[11]. Derlemi oluşturan yazılımlar Linux tabanlı bir sunucu üstünde Debian 6 (Squezee x86_64) işletim sistemi ile çalışmaktadır.

Derlem altyapısında CWB/CQPWeb altyapısı seçilmesinin temel nedenlerinden biri de bu çatının esnekliği ve çok büyük sayıdaki sözcüklerden oluşan derlemleri oldukça hızlı şekilde işleyebilmesidir. CQP aramaları önbelleğe alabilmekte, tekrar edilen sorguları önbellekten (cache) çağırabilmektedir. TS Corpus üstünde 8 milyon 910 bin 6 kere bulunan "ve" sözcüğü, toplam 491 milyon 360 bin 398 sözcük içinden ilk aramada 6.4 saniyede sorgulanırken, ikinci aramayı önbellekten çağırarak sorguyu 0.486 saniyede tamamlamaktadır. Benzer şekilde düzenli ifade kullanımında da önbellekleme işlevi çalışmaktadır. Düzenli ifade ile yapılan *yüz** sorgusu ilk aramada 9.4 saniye içinde 1 milyon 188bin 962 sonuç getirirken ikinci arama önbellekten sadece 0.041 saniyede getirilebilmektedir.

5. TS Corpus'un Kullanımı

5.1. Aramalar

TS Corpus kullanıcı dostu bir arayüz ile sunulmaktadır. Derleme erişmek için, derlem internet sitesinde bulunan kayıt formunu doldurarak kullanıcı adı ve parola oluşturmak gereklidir. CQPWeb altyapısında olmayan otomatik kullanıcı oluşturma-yetkilendirme özelliği, CQPWeb yetkilendirme ve kullanıcı yapısının mod_auth_myqsl[12] Apache modülü ile CQPWeb altyapısına eklenmesiyle sağlanmıştır. Derleme üç ayrı alan adından (domain) ve bir altalan (subdomain) adından erişmek mümkündür[13].

TS Corpus üstünde yapılacak aramalar iki ayrı arama yöntemiyle yapılmaktadır. Öntanımlı yöntem Basit Arama'dır (Simple Query). Basit aramalar büyük-küçük harf duyarlı olarak ve olmayarak iki şekilde yapılabilir. İkinci arama yöntemi ise CQP Sözdizimi (CQP Syntax) aramalarıdır. Bu aramalar CQP altyapısının kendi dili kullanılarak yapılır ve CQP'nin sağladığı hemen hemen bütün arama yeteneklerini grafik arayüz üstünden kullanmayı sağlar. CQP'nin arayüzü üzerinden kullanıcılara sunulmayan, ancak komut satırı üstünde kullanılabilen bir arama altyapısı da vardır. TS Corpus 4 ayrı katmanda etiketlenmiştir. Bu katmanlar ve bu katmanlar üstünde yapılabilecek aramalar aşağıda verilmiştir.

Sözcük (Word): Öntanımlı olarak derlem arayüzünde basit aramalarda kullanılan katmandır. Aranan sözcük sorgulandığı şekilde veya düzenli ifadeler (regular expressions) kullanılarak bu katmanda sorgulanır.

gelebilir	Basit Arama
[word="gelebilir"]	CQP Sözdizimi Araması

Bu katmanda joker karakterler ve düzenli ifadeler de kullanılabilir. Kullanılabilen joker karakterler **? * + , : @ / ( ) [ ] { } _** - < > karakterleridir.

8 TS Corpus sürüm 2 için CWB sürüm 3.4.7 ve CQPWeb sürüm 3.0.9
9 sürüm 2.2.16
10 PHP 5.3.3-7+squeeze15 with Suhosin-Patch
11 sürüm 14.14, build 5.1.66
12 http://modauthmysql.sourceforge.net/
13 http://tscorpus.com http://turkishcorpus.com http://türkcederlem.com http://gui.tscorpus.com

gel*	Gelebilir, gelir, gelincik vb.
g?l	Gül, gel, gol vb.
k[a,e,u,ı]l?	Kale, kulu,kalp, kala,kalk,kelt vb.

Sözcük Türü Etiketi (PosTag): Sözcük türü etiketinin bulunduğu katmandır. Basit aramalarda, aranan sözcük türü etiketinin önüne alt çizgi (under score) yazılarak arama yapılır. CQP sözdizimi ile yapılan aramalarda sorgu [PosTag=”ETİKET”] şeklinde oluşturulur.

_NOUN	Basit Arama
[PosTag=”NOUN”]	CQP Sözdizimi Araması

Bu arama ile TS Corpus üstünde işaretlenmiş sözcük türü etiketleri sorgulanabilir. Sorgular tek bir etiket içerebileceği gibi aynı sorguda birden fazla etiket ile ardışık yapılar da sorgulanabilir. Örneğin *_Adj* sorgusu ile derlemdeki tüm adjective (sıfat) olarak işaretlenmiş sözcüklere ulaşılabilirken *_Adj _Conj _Adj* sorgusu sıfat+bağlaç+sıfat dizilimindeki sonuçları getirecektir. Aynı zamanda sorguda etiket yerine sözcük de kullanılabilir. *_Adj ama _Adj* sorgusu bir sıfat, ardından gelen “ama” bağlacı ve yine bir sıfat dizilimindeki (nazik ama ince vb.) sonuçları döndürecektir. Sözcük türü aramalarını belirli bir sözcüğün sadece istenen sözcük türündeki kullanımlarına ulaşmak için de kullanabilir.

Örneğin “gül” kökünün yalnızca Verb (eylem) olarak işaretlendiği durumları aramak için {gül*}_Verb sorgusu, isim olarak işaretlendiği durumları aramak için {gül*}_Noun sorgusu kullanılır.

{gül*}_Verb	gülerdim,gülmekten, gülüp vb.
{gül*}_Noun	Gültepe, güller, Gülhane vb.

Kök (Lemma): Derlem arayüzünde görüntülenen sözcüğün biçimbirimsel çözümleyici tarafından bulunan kök halini sorgulamak için kullanılır. Basit aramalarda sorgu {ARANAN_KÖK}, CQP sözdiziminde ise [LEMMA=”ARANAN_KÖK”] olarak oluşturulur.

{burun}	Basit Arama
[LEMMA=”burun”]	CQP Sözdizimi Araması

Bu arama ile ses düşmesine veya değişimine uğrayan kökleri bulmak da mümkündür. Örneğin “burun” sözcüğünü sorguladığınızda burnum, burnun, “kitap” sözcüğünü sorguladığınızda kitaplık, kitabım sözcükleri sonuçlar içinde görüntülenecektir.

Biçimbirim (Morph): Derlemdeki tüm birimlerin biçimbirimsel etiketlerinin sorgulaması bu katmanda yapılır. Basit aramalar bu seviyede kullanılmaz. CQP sözdizimi ile yapılacak aramalarda sorgu [Morph=”ARANAN_BİÇİMBİRİMSEL_ETİKET”] şeklinde oluşturulur.

----------	Basit arama bu seviyede kullanılmaz
[Morph=".*\+Loc\+.*"]	CQP Sözdzimi Araması

Bu katmanlar üzerinde yapılacak aramaların yanısıra CQP altyapısı birleşik aramalar yapmaya da izin vermektedir. Birleşik aramalar ile ardışık sözcükleri, belirli bir bağlam sınırı içinde sorgulamak da mümkündür.

Örneğin kullanıcı, CQP sözdizimi araması ile

[Lemma="televizyon"] + [Lemma="izle" | Lemma="seyret" | Lemma="bak"]
sorgusunu yaptığında, “televizyon” sözcüğünün derlem içindeki her türlü görünümü ile bunu takip eden, izle, seyret ve bak sözcüklerinin her türlü görünümüne tek arama ile ulaşabilir. Sonuç seti bu üçlü veriyi içerecek şekilde oluşturulacaktır.

5.2. İşlemler

Yukarıda örneklenen sorgular aramaların ilk adımını oluşturmaktadır. Arama tamamlandığında derlem arayüzü ana sonuç sayfasına otomatik olarak yönlendirilmektedir. Ana sonuç sayfasında yapılan sorgu, toplam sonuç sayısı, derlemdeki toplam sözcük sayısı ve yapılan işlemin tamalanma süresi yer almaktadır.

Ana sonuç sayfasında getirelen sonuçlar öntanımlı olarak bağlam içinde anahtar sözcük/KWIC[14] şeklinde gösterilmektedir. Bulunan anahtar sözcük üstüne imleç getirildiğinde sözcük türü etiketlerini barındıran bir bilgi ekranı açılarak anahtar sözcüğün sözcük türü etiketini göstermektedir. Bulunan sonuç üstüne tıklandığında, kullanıcı sözcüğün geçtiği metni içeren yeni bir pencereye yönlendirilmektedir. Bu pencerede bulunan menü ile (varsa) metine ait üst metin bilgisine ulaşılabildiği gibi, anahtar sözcüğün bulunduğu metnin daha geniş bir örneklemi de çağırılabilir. Ayrıca yine bu pencerede bulunan “etiketleri göster” (Show Tags) düğmesine tıklanarak tüm sözcüklerin biçimbirimsel çözümlemesi de çağırılabilir. Bunun yanı sıra ana sonuç sayfası işlemler (actions) menüsünü de barındırmaktadır. Bu menü “yeni arama, aramayı limitleme, sıklık dökümü, sıralama, sonuçları indirme, kategorize etme ve sonuçları kaydetme” işlemlerini içerir.

Yeni Arama (New Query)

Kullanıcıyı yeni arama yapmak üzere derlem ana sayfasına yönlendirir.

Aramayı Limitleme (Thin)

Derlemde yapılan bazı sorgular milyonlarca sonuç döndürebilmektedir. Aramayı limitleme seçeneği ile toplam sonuç içinden bir örneklem almak mümkündür. Bu örneklem sonuç seti içinden rastgele seçilecektir. Ayrıca istenen örneklem yeniden üretilebilir (selection is reproducible) veya yeniden üretilemez (selection is not reproducible) olarak seçilebilir.

Sıklık Dökümü (Frequency Breakdown)

Sıklık dökümü bulunan sonuçlar içinde her bir sonucun kaç defa derlem içinde bulunduğunu göstermek ve bulunan oranı yüzde olarak hesaplamak için kullanılır. Sıklık dökümü sonuçları ayrıca sözcük türü etiketlerini içerecek şekilde de oluşturulabilir. Tüm sıklık listelerini kullanıcı kendi bilgisayarına indirilebilir.

Dağılım (Distribution)

Dağılım özelliği sorgulanan girdinin derlemi oluşturan metin türleri, alanlar vb. üst metin bilgisi içindeki rakamsal dağılımını, toplam sözcük sayısına ve alanlara olan oranını göstermek için kullanılır. TS Corpus içinde yalnız iki ayrı veri kümesi (gencor ve newscor bkz. 3.Veri) bulunmaktadır. Dolayısıyla bu özellik şu anda yapısal olarak elverişli olmasına rağmen etkili bir şekilde kullanılamamaktadır. Dağılım tablo halinde ve grafik olarak gösterilmektedir.

Sıralama (Sort)

Sıralama özelliği iki aşamada kullanılır. İlk aşamada bulunan sonuç kümesi içindeki anahtar sözcük/yapı alfabetik olarak sıralanmaktadır. Bu işlem kullanıcıyı otomatik olarak yeni bir pencereye taşır. Bu pencerede öntanımlı olarak anahtar sözcüğün alfabetik olarak sıralandığı sonuçlar görüntülenecektir. Bu penceredeki seçenekler kullanılarak sıralamanın anahtar sözcüğün solundan veya sağından kaçıncı sözcüğe göre yapılabileceği seçilebilir. Ayrıca bu sıralamada istenen konumda bulunacak sözcük türü etiketini seçmek de mümkündür. Örnek bir aramayı şu şekilde yapabiliriz. Kullanıcı, derlem ana ekranından *şeytan** sorgusunu basit arama ile sorguladığında arama yapılacak ve kullanıcı otomatik olarak sonuç ekranına yönlendirilecektir. Bu ekrandaki işlemler menüsünden sıralama (sort)

14 KWIC – KeyWord In Context

komutu veren kullanıcı ilgili ekrana yönlendirilecek ve anahtar sözcükler alfabetik olarak sıralanmış olacaktır. Bu ekranda kullanıcı konumu anahtar sözcüğün hemen öncesi (1 Left) ve etiket sınırlamasını (Tag Resriction) _Adj olarak seçtiğinde herhangi bir sıfatı takip eden şeytan sözcüğünün derlem içindeki tüm görünümlerine ulaşmış olacaktır. (acımasız şeytan, aksi şeytan, yeni şeytanlıklar vb.)

Eş dizim Örüntüleri (Collocations)

TS Corpus eşdizim örüntülerine ulaşmayı da sağlamaktadır. Ana sonuç ekranında bulunan işlemler menüsü kullanılarak eşdizim örüntülerine ulaşılabilir.

Sonuçları İndirme (Downloads)

Kullanıcılar arama sonuçlarını kendi bilgisayarlarına indirerek saklayabilir, kullanabilir. CWB altyapısı bu sonuçların düz metin belgesi yınasıra sıklıkla kullanılan ofis yazılımları ve Filemaker gibi veri tabanı yazılımlarıyla uyumlu olarak kaydedilmesine de izin vermektedir.

Kategorilere Ayırma (Categorization)

Ana sonuç ekranındaki işlemler menüsünde bulunan kategorilere ayırma özgün bir işlevdir. Sonuçların, "kullanıcının kendi belirlediği kategori başlıklarına" göre işaretlenmesini ve bu şekilde saklanarak istendiğinde geri çağırılmasını sağlar.

Sonuçları Kaydetme (Save Current Set of Hits)

Kullanıcılar yaptıkları sorgunun sonuçlarını derlemin bulunduğu sunucu üstünde kaydedebilir.

5.3. Diğer Özellikler

Derlem ana sayfasında bulunan aramalar (Corpus Queries) ve Kullanıcı Kontrolleri (User Controls) menüleri derlemin sunduğu farklı özelliklere kullanıcının ulaşmasını sağlamaktadır.

Kullanıcı Kontrolleri (User Controls)

Kullanıcı ayarları (User Settings), Arama Geçmişi (Query History), Kaydedilmiş Aramalar (Saved Queries), Arama Yükleme (Upload a Query) ve Altderlem Oluştur/Düzenle (Create/Edit Subcorpora) başlıklarını içeren bu menü ile kullanıcılar kendi tercihlerini ayarlayabilir, arama geçmişini görebilir, önceden kaydettiği aramalara ulaşabilir, daha önce bilgisayarlarına indirdiği bir aramayı çalışmak üzere derleme yükleyebilir ve alt derlemler oluşturabilir, aynı zamanda altderlemleri düzenleyebilirler.

Aramalar (Corpus Queries)

Bu menü Standart arama (Standard Query), Sınırlandırılmış Arama (Restricted Query), Sözcük Arama (Word Lookup), Sıklık Listeleri (Frequency Lists) ve Anahtar Sözcükler (Keywords) başlıklarını içermektedir. Kullanıcılar aramalarını derlemin tamamında veya bir kısmında yapmak için standart arama veya sınırlandırılmış arama seçeneklerinden birini seçebilirler.

Sözcük arama özelliği düzenli ifadeleri de içerir şekilde bir sözcüğün veya yapının bulunduğu konumla sınırlandırılarak aranabilmesini sağlar. Sorgular "ile başlıyor", "ile bitiyor", "içeriyor" ve "tam eşleşme" konumlarından biri seçilerek yapılır. Sonuçlar sadece sözcük veya sözcük ve sözcük türü etiketini içerecek şekilde çağırılabilir.

Sıklık listesi özelliği ile derlemde bulunan katmanlardan herhangi biri ile arama yapılabilir. Sorgular, sözcük arama özelliğinde olduğu gibi "ile başlıyor", "ile bitiyor", "içeriyor" ve "tam eşleşme" konumlarından birini içermektedir. Sonuçların hangi sıklık aralığında olabileceği kullanıcı tarafından ayarlanabilir.

Anahtar sözcükler ise, eğer CWB/CQPWeb altyapısı birden fazla derlem barındırıyorsa kullanılabilmektedir[15]. Derlem yöneticisi tarafından kullanıcıların erişimine açılmış sıklık listeleri arasındaki pozitif/negatif anahtar sözcükler bu menüden sorgulanmaktadır.

15 30 Ağustos 2013 taririhden itibaren TS Corpus iki ayrı derlemi barındırmaktadır. TS Corpus sürüm 2 ve TS Corpus Wikipedia -Beta- derlemlerinin sıklık listeleri karşılaştırılabilir durumdadır.

6. Sonuçlar

TS Corpus projesi belirli bir veri setini kullanarak tek bir derlem oluşturma çalışması değil, Türkçe derlem dilbilim çalışmalarına katkıda bulunmak üzere farklı derlemler oluşturmayı ve bu amaçla derlem oluşturmada kullanılabilecek araçlar/betikler üretmeyi hedefleyen bir çalışmadır.

Bu çalışmanın yapıldığı tarihte TS Corpus projesi iki ayrı derlemi barındırmaktadır.[16] Bu derlemler sözcük türü ve biçimbirimsel olarak etiketlenmiş toplam 537M+ sözcüklük bir veriyi kullanıcılara sunmaktadır. Türkçe için ilk defa yapılmış bir çalışma olarak bu kadar büyük veri setlerinin sıklık listeleri birbiriyle karşılaştırılabilir, farklı derlemlerdeki pozitif/negatif anahtar sözcükler bulunabilir durumdadır.

Bu noktada çalışmanın sağladığı önemli katkılardan biri de büyük veri setlerinden kullanıcıların faydalanmasını sağlamış olmaktır. TS Corpus sürüm 2'ye kaynak olan BOUN Web Corpus, 4.4 Gb ham veri büyüklüğü ile düzenli ifade kitaplıklarını kullanamayan, metin editörleri (Vi, eMacs vb.) ile veri üstünde çalışma becerisine sahip olmayan kullanıcılara da bu veriyi erişilebilir bir kaynak olarak kullanma olanağı sağlamıştır. Kelime işlemci yazılımlar kullanarak, grafik arayüz üstünden bu veriyi açmak, üstünde ara-bul gibi basit işlemler yapmak mümkün değildir[17]. (TS Corpus sürüm 2'nin işaretlenmiş, CWB/CQPweb altyapısına gönderilmeye uygun biçiminin dosya boyutu 15.5GB, TS Corpus Wikipedia -Beta-'nın dosya boyutu 1.5GB'dir.)

Bunun ötesinde Türkçe için ilk defa kullanıcıların sözcük türü ve biçimbirimsel etiketler ile arama yapması sağlanmıştır. Anahtar sözcüğün sağında ve solundaki belirli bir aralıktan sözcük türü etiketleri ile arama yapılması da çalışmanın katkıları arasındadır. TS Corpus verisi kullanılarak elde edilmiş sıklık listeleri derlem internet sitesinde yayınlanmıştır. Bu listeler “sözcük türü” ve ”biçimbirimsel etiket” içeren Türkçe sıklık listeleridir[18].

Çalışma, açık kaynak kodlu yazılımlar ve halihazırda kullanılabilir şekilde ancak dağınık halde bulunan yazılımlar ile yapılabileceklerin bir örneğini göstermesi açısından da önemlidir.

CWB/CQPweb altyapısının Türkçe bir derlem ile uyumlu çalıştığı görülmüştür. TS Corpus bu altyapıyı kullanan ilk Türkçe derlemdir. Farklı dillerde ve çok büyük miktarda veri içeren derlemler CWB/CQPWeb altyapısıyla kullanıcılara sunulmaktadır[19]. Bu yapının sağladığı arama hızı ve özellikler benzer diğer derlem arayüzü-altyapı yazılımlarına oranla daha gelişmiştir.

Bu, derlemdilbilim kuramları çerçevesinde hazırlanacak dili temsil yeteneğine sahip ve dengeli bir Türkçe derlemin aynı zamanda sözcük türü, biçimbirim ve hatta yapısal olarak işaretlenmiş bir derlem olarak, CWB/CQPWeb altyapısının sağladığı bütün özellikleri kullanarak oluşturulabileceğini göstermektedir. Türkçe derlem oluşturma sırasında yeni bir derlem arayüzü oluşturmaya harcanacak bütçe, zaman ve işgücünden tasarruf etmek bu yolla mümkündür.

TS Corpus projesi Türkçe derlem dilibilim alanındaki bazı eksikliklerin de daha net şekilde ortaya serilmesine katkıda bulunmuştur. İşaretlenmiş Türkçe derlemler oluşturmak amacıyla başarım oranı daha yüksek bir sözcük türü etiketleme yazılımına ihtiyaç vardır. Bu yazılımların eğitilebileceği (train) ve test edilebileceği derlem ihtiyacı, TS Corpus gibi büyük veri içeren derlemler kullanılarak karşılanabilir.

TS Corpus Wikipedia -Beta- derlemiise Türkçe Wikipedia sayfalarını CWB/CQPWeb altyapısıyla sunmaktadır. Bu önsürüm, Wikipedia maddelerinin makine öğrenme (machine learning) yöntemleriyle ayrıştırılmasından sonra tamamlanmış olacaktır.

TS Corpus sürüm 3 için de çalışmalar devam etmektedir. Bu sürüm internetten toplanmış ve üst metin bilgilerine göre sınırlandırılmış aramalar yapılacak şekilde düzenleniş, yine sözcük türü ve biçimbirimsel olarak işaretlenmiş, 500 milyon– 1 milyar arası sözcükten oluşması hedeflenen bir çalışmadır.

Bu çalışmanın yapılabilmesi için veri toplayıcı (crawler) yazılımları derlem sunucuları üstünde test edilmektedir. Çalışmanın yapıldığı dönemde iki ayrı crawler ile 1.7 milyon internet sayfası indexlenmiştir. Bu çalışmanın bir uzantısı olarak TS Corpus arama motoru kullanıcıların erişime açılmıştır[20].

Bu verinin Makine öğrenme yöntemleriyle sınıflandırma (classification) ve kümeleme (clustering) yapılarak üstmetin etiketleriyle işaretlenmesi hedeflenmektedir. Bu konuda yapılacak çalışmalar her gün büyüyen internet üstündeki verinin dilbilim çalışmalarında kullanılmasını sağlayacak şekilde otomatik olarak bilgisayarlara işletilmesi anlamında büyük önem taşımaktadır.

---

16 TS Corpus sürüm 2 ve TS Corpus Wikipedia -Beta-

17 Ülkemizde sıklıkla kullanılan Microsoft Word yazılımının çalıştırabildiği maximum boyutu 512 Mb'dir. http://support.microsoft.com/kb/2599449 & http://support.microsoft.com/kb/211489

18 http://www.tscorpus.com/en/frequency-english

19 http://cwb.sourceforge.net/demos.php

20 TS Corpus arama motoru TS Corpus sürüm 3'ü oluşturması planlanan veri içinde arama yapmaktadır. Arama sayfası 1 Mayıs 2013 tarihinde erişime açılmıştır. http://gui.tscorpus.com/search/ adresinden arama motoruna erişmek mümkündür. İkinci bir arama motorunun önümüzdeki dönemde erişime açılması planlanmaktadır.

TS Corpus çalışması ayrıca günümüz teknolojisine uygun şekilde tablet bilgisayalar ve akıllı telefon gibi cihazlarla uyumlu şekilde CQPWeb arayüzünün geliştirilmesi konusunda da çalışmaktadır. Bu bağlamda oluşturulan yeni arayüzün erken bir önsürümü derlem ana sayfasında bulunan bağlantı ile denenebilir durumdadır[21].

Ekler

Ek 1: TS Corpus Sözcük Türü Etiketleri

#	PosTag Etiketi*	Sözük Türü
1	_Noun	Ad
2	_Verb	Eylem
3	_Adj	Sıfat
4	_Adv	Zarf
5	_Conj	Bağlaç
6	_Det	Belirteç
7	_Dup	İkileme
8	_Interj	Ünlem
9	_Ques	Soru
10	_PostP	Edat
11	_Num	Numara
12	_Pron	Zamir
13	_Punc	Noktalama
14	_UnDef	Belirsiz

21 TS Corpus Beta Interface olarak adlandırılan bu arayüz CQPWeb'ün öntanımlı arayüzünde bulunan CSS ve HTML yapısının değiştirilmesi/geliştirilmesi ile oluştulmuştur. Arayüz derleme yeni bir özellik ekleme arayışında değildir.

*Sorgularda kullanılan PosTag etiketleri büyük/küçük harf duyarlıdır.

Kaynakça

Atkins, S., Jeremy, C. and Nicholas, O. (1992) *Corpus design criteria*, Literary and Linguistic Computing 7:1-16.

Biber, Douglas, Susan Conrad and Randi Reppen. (1998) *Corpus Linguistics*, Cambridge University Press.

BOUN Web Corpus. 28 Nisan 2013 tarihinde http://79.123.177.209/~hasim/langres/BounWebCorpus.tgz adresinden erişildi.

Evert, S., The OCWB Development Team (2010) *The IMS Open Corpus Workbench (CWB) CQP Corpus Encoding Tutorial*. http://cwb.sourceforge.net

IMS CWB. 28 Nisan 2013 tarihinde http://cwb.sourceforge.net/demos.php adresinden erişildi.

McEnery, T., Baker, P. and Hardie, A. (2006) *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press

Microsoft Support. 24 Nisan 2013 tarihinde http://support.microsoft.com/kb/2599449 adresinden erişildi.

Microsoft Support. 24 Nisan 2013 tarihinde http://support.microsoft.com/kb/211489 adresinden erişildi.

Sak, H., Güngör, T. and Saraçlar, (2007) *Morphological disambiguation of Turkish text with perceptron algorithm*. In *CICLing 2007*, LNCS 4394, p:107-118

Sak, H., Tunga, G., Saraçlar, M.. (2008) *Turkish Language Resources: Morphological Parser, Morphological Disambiguator and Web Corpus.* GoTAL 2008: p417-427.

Sak, H., Güngör, T. and Saraçlar, (2011) Resources for Turkish morphological processing. *Language Resources and Evaluation*, 45(2)p:249-261

Haşim Sak. 24 Nisan 2013 tarihinde http://www.cmpe.boun.edu.tr/~hasim/ adresinden erişildi.

Sezer, Taner. (2010) *Corpus Linguistics Theory and Design and Application of a Turkish Corpus*. Unpublished MA Thesis. Mersin

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

TS Corpus – Frequency Lists. 12 Şubat 2013 tarihinde http://www.tscorpus.com/en/frequency-english adresinden erişildi.

TS Corpus Search Engine. 26 Nisan 2013 tarihnde http://gui.tscorpus.com/search/ adresinden erişildi.