Web Tabanlı Türkçe Ulusal Derlemi (TUD)

Yeşim Aksan[1], Mustafa Aksan[1], Selma Ayşe Özel[2], Hakan Yılmazer[3],

Umut Ufuk Demirhan[1], Ümit Mersinli[1], Yasin Bektaş[4], Serap Altunay[1]

[1] Mersin Üniversitesi, İngiliz Dili ve Edebiyatı Bölümü, Mersin

[2] Çukurova Üniversitesi, Bilgisayar Mühendisliği Bölümü, Adana

[3] Mersin Üniversitesi, Bilgi İşlem Daire Başkanlığı, Mersin

[4] Mersin Üniversitesi, Erdemli Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Mersin

yesim.aksan@gmail.com, mustaksan@gmail.com, saozel@gmail.com, yilmazerhakan@gmail.com, umutufuk@gmail.com, umit@mersinli.org, ybektas79@gmail.com, serapaltunayy@gmail.com

Teşekkür: TUD TÜBİTAK 108K242 (2008-2011) tarafından desteklenmiştir.

Özet: Bu çalışma, Türkçe'nin ilk kapsamlı ve genel amaçlı derlemi olan Türkçe Ulusal Derlemi (TUD)'ni [1] tanıtmayı amaçlamaktadır. Tanıtım Sürümünü Ekim 2012'de yapan TUD'un derlem dilbilim ilkeleri temelindeki derlem tasarım ölçütleri, web tabanlı ve kullanıcı dostu arayüz yazılım mimarisi özellikleri açıklandıktan sonra, kullanıcıların TUD üzerinde gerçekleştirdikleri sorgulamalarda yararlanacakları derlem araçları listelenecektir. Son olarak, dili temsil etme yeterliliğine sahip TUD gibi dil kaynaklarının sosyal bilimler ve bilgisayar mühendisliği alanlarındaki araştırmalardaki yeri ve önemi üzerinde durulacaktır.

Anahtar Sözcükler: Derlem dilbilimi, Derlem Tasarımı, Web-tabanlı derlem arayüzü, Türkçe Ulusal Derlemi (TUD)

Web-Based Turkish National Copus (TNC)

Abstract: The aim of this study is to describe the construction process of the first large scale, general-purpose corpus of Turkish, namely Turkish National Corpus (TNC). The paper is organized as follows; first, the overall design features of TNC-Demo – released in October 2012 – which are based on principles well-defined in corpus linguistics, will be presented. Secondly, the web based interface architecture and pre-defined functions and tools of the TNC interface that will aid users in making their queries will be shown. Finally, the role of representative language resources like TNC in disciplines of social sciences and computer engineering is discussed.

Keywords: Corpus linguistics, Corpus building, web-based corpus interface, Turkish National Corpus (TNC).

1. Giriş

Bir dil kaynağı olarak derlem, belli amaçlar temelinde yapılandırılmış metinler/konuşmalar bütünüdür. Genel amaçlı (İng. reference/general) hazırlanan bir derlemi şöyle tanımlayabiliriz. Belli bir dili temsil edebilme amacıyla, belli bir zaman aralığında, yazılı ve/veya sözlü dil kullanım metinlerini/konuşmalarını, yazar/konuşan özelliklerini (cinsiyet, yaş, eğitim vb.), iletişim ortamlarının alan ve türlerine (İng. domain, genre) ve yayın ortamlarına (kitap, süreli yayın vb.) göre dengeli ve katmanlı örnekleme yoluyla derleyip, belirlediği

ölçütleri kapsayan ayrıntılı veribilgisi (İng. metadata) ve temel dilbilimsel çözümleme araçlarıyla birlikte elektronik ortamlarda sunan kaynaklara derlem denir [22; 15]. Bilgisayar teknolojilerindeki hızlı gelişmeler sayesinde gerçek dil kullanımını içeren büyük derlemler oluşturulabilmiştir (örn., British National Corpus [4], Corpus of Contemporary American English). Bu derlemler üzerinden yürütülen dilbilim ve bilişim alanındaki çalışmalar ile dilin başka yöntem ve araçlarla görülemeyen pek çok önemli özelliği ortaya çıkarılmıştır. Günümüzde çok sayıda dilin özel ya da genel amaçlı derlemleri kurulmuş ve kullanıcıların hizmetine sunulmuştur [14].

Türkçe için yukarıda sunulan dil derlemi tanımına en yakın derlem ODTÜ Türkçe Derlemi'dir [21]. 1990 sonrası yazılı metin örneklerini 291 farklı veri kaynağından alan ve 2 milyon sözcükten oluşan Türkçenin bu ilk yazılı derlemi, on farklı metin türünü kapsamaktadır. Kullanıcılar derlemi çevrimdışı, platform bağımlı bir yazılım aracılığıyla çalıştırabilmektedir. Söz konusu yazılım basit ve düzenli ifade sorgusu yapmaya olanak tanırken, dil derlemi arayüz özelliklerinden olan bağımlı dizin satırları (İng. concordance lines), listeleme (İng. sorting), dağılım (İng. distribution) ve sayısal sıralı eşdizimlilik listeleri (İng. collocation lists) gibi derlem araçlarına sahip değildir. Son 15 yılda internette birçok dile ilişkin verinin yer alması www'i hızlı, kolay ve insan gücü gerektirmeksizin, otomatik biçimde dil derlemi kurmak için kullanılır yapmıştır [11]. Bu yöntemle Türkçe için geliştirilen TurkishWaC [2], kaynak sözcük tarama (İng. seed word) yoluyla Wikipedia sayfalarından edinilen 42 milyon sözcükten oluşmaktadır. Bu derlem dilbilim ve sosyal bilimler alanlarında sözcük profili çalışması yapmaya uygun, ücretli Sketch Engine (http://sketchengine.co.uk) derlem sorgulama sistemiyle kullanıcıların erişimine açıktır. Bunun yanı sıra, Türkçe için bilgisayar mühendisleri tarafından hazırlanan yazılı Türkçenin derlemleri, verilerini yine www'den almış, sözcük sayısı açısından büyük ancak derlem tasarımı ilkelerine uymayan derlemlerdir. Bu derlemlerin birçoğu bilgisayar mühendislerinin Türkçe için geliştirdiği yazılımları sınamak ve Türkçenin sözcük, tümce, ek vb. dilsel birimlerinin nicel dökümünü almak üzere derlenmiş metinler bütündür. TurCo [8] on farklı internet sayfası kaynak alınarak bir araya getirilmiş ve 44 milyon sözcük içeren bir derlemdir ve bu derlem kullanılarak Türkçe sözcüklerin ve sözcük takımlarının bazı istatistiksel özellikleri saptanmıştır. BOUN Derlemi [20] Türkiye'de okunan başlıca üç farklı gazetenin internet sayfalarını içeren dört farklı alt derlemi kapsamakta ve 423 milyon sözcükten oluşmaktadır. Bu derlem üzerinden Türkçe sözcüklerdeki biçimbirimlerin istatistiksel bir modeli geliştirilmiştir. BOUN Derlemi XML formatında araştırmacıların ulaşabileceği bir dil kaynağıdır. Son olarak bu grup içinde, Türki diller arasında biçimbirimsel çözümleme yazılımlarını sınamak için kurulan 3.37 milyar büyüklüğünde [5], araştırmacıların ulaşamadığı Türkçe derlem bulunmaktadır.

Bu yazıda tanıtmayı amaçladığımız Türkçe Ulusal Derlemi (TUD) yukarıda özetlenen derlemlerden farklı olarak, derlem dilbilimin derlem kurma ilkelerine göre geliştirilen, en iyi uygulamaları örnek alan ve derlem tasarım sürecine uyarlayan, web tabanlı ve kendine özgü arayüzü olan, dili temsil gücüne (İng. representativeness) sahip, dengeli (İng. balanced), yazılı ve sözlü Türkçe metin örneklerini içeren Türkçenin ilk referans derlemidir.

2. Türkçe Ulusal Derlemi tasarım ölçütleri

Derlem tasarımı temel olarak beş ilkeden oluşur. Derlemin temsil gücü, denge, örneklem, zaman içindeki değişim ve derlem metinlerini belirleme bir derlem oluştururken dikkat edilmesi gereken ilkelerdir [25]. Derlemin temsil gücü, derlemi oluşturan örneklemin dil değişkelerini ne ölçüde kapsadığını gösterir [6]. Denge, derlemi oluşturacak türlerin kapsamını belirtmektedir. Bir derlem tasarlanırken olabildiğince geniş metin türlerini içermesi

hedeflenmelidir ancak, derlem dengesi için bilimsel bir ölçüt bulunmamaktadır. Derlemleri oluşturan araştırmacılar genellikle daha önce yapılmış olan bir derlemi kendilerine model olarak alırlar. Örneklem ise, her tür için metin parça/bütün seçimini; zaman içindeki değişimi, derlemi durağan (İng. static) ya da dinamik (İng. dynamic) bir dil modeli olarak ele almayı gösterir.

TUD tasarım ilkeleri British National Corpus (BNC) [4] örnek alınarak geliştirilmiştir. Eşzamanlı, durağan bir derlem olarak tasarlanan TUD, 50 milyon sözcükten oluşan, 20 yıllık bir dönemi kapsayan, günümüz Türkçesinin çok sayıda farklı konu alanı ve metin türünden yazılı ve sözlü örneklerini içeren (%98'i yazılı %2'si çeviriyazıya geçmiş sözlü dil verisi) geniş kapsamlı bir referans derlemdir. Derlemin yazılı metin örneklerini içeren ve Ekim 2012'de kullanıcıların erişimine açılan TUD-Tanıtım Sürümü 1990-2009 yılları arasında yayımlanan yazılı ve sözlü toplamda 4442 veri kaynağından seçilen, 9 konu alanını ve 39 dilsel türü (bilimsel makaleler, roman, e-postalar, bloglar vb.) içeren metin örneklerinden oluşmaktadır (bkz. Tablo1).

TUD'un derlem metinleri ya da metin parçaları dil dışı ölçütlere göre belirlenmiştir. Bunlar, metinlerin konu alanı, metinlerin yayınlanma tarihi ve yayın ortamıdır. Konu alanı kurgusal ve bilgilendirici metinlerden oluşmaktadır. Yazınsal metinler (roman, kısa öykü, şiir, tiyatro) kurgusal alanı temsil etmektedir. Toplumbilimleri, sanat, ticaret-finans, düşünce-inanç, dünya sorunları, uygulamalı bilimler, doğa-temel bilimleri, sanat, hobi, yemek tarifi gibi serbest olarak adlandırılan metinler ise bilgilendirici alan için örneklem almak üzere seçilmiştir. Yayın ortamı olarak kitaplar, süreli yayınlar (gazete, dergi), çeşitli (yayınlanmış-yayınlanmamış) metinler ve konuşmak üzere yazılmış metinler kullanılmıştır.

Alan	Oran	Toplam Sözcük Sayısı
1. Dünya Sorunları	% 20,05	9.591.797
2. Kurgusal Düzyazı	% 19,22	9.194.674
3. Serbest	% 14,96	7.155.998
4. Toplum Bilimleri	% 14,55	6.961.521
5. Ticaret ve Finans	% 9,21	4.404.453
6. Sanat	% 7,50	3.586.866
7. Uygulamalı Bilimler	% 7,19	3.441.050
8. Düşünce ve İnanç	% 4,31	2.061.068
9. Doğa ve Temel Bilimler	% 2,96	1.419.861
TOPLAM	% 100	47.817.288

Tablo 1. TUD-Tanıtım Sürümünde Metinlerin Konu Alanlarına göre Dağılımı

3. TUD-Tanıtım Sürümü Yazılım Mimarisi

3.1. Genel Özellikler

TUD-Tanıtım Sürümü 4 çekirdekten oluşan, 3.20GHz hızında, 8MB önbellekli, 1 adet Intel® Xeon® E3-1225v2 işlemcili; 16 GB bellek ve 1 TB sabit disk alanına sahip; FreeBSD 9.0 [23] işletim sistemini kullanan bir sunucu üzerinde bulunmaktadır. Derlem metinleri ve dizin yapısı MySQL 5.5.22 [16] veritabanı yönetim sisteminde oluşturulmuş bir veritabanında yer almaktadır. TUD-Tanıtım Sürümü web tabanlı olup, web arayüzü aracığıyla kullanım ve sorgulama imkânı sunmaktadır. Web arayüzü açık kaynaklı kodlar kullanılarak hazırlanmıştır. Bu amaçla web sunucusu olarak Apache/2.2.22 (FreeBSD) [3] kullanılmış olup, kullanıcı ara yüzleri PHP 5.4.21 [17], HTML [10], CSS [7], Javascript [12], Jquery [13] ile hazırlanmıştır. Ham metinleri işleyip, sözcükbirimlerin (İng. token) ve teksözcüklerin (İng. type) çıkarılmasında Perl 5.12.4 [24] betik dili kullanılmıştır.

Sunucu işletim sisteminin, UNIX tabanlı ve açık kaynak kodlu olması ileri seviyede ağ, performans, güvenlik ve uyumluluk özelliklerini beraberinde getirmiş; bunun yanı sıra sunucu uygulamalarının ve modüllerinin uygulanabilirliği açısından gelişmiş port yapısı ile esnek bir çalışma ortamı sağlamıştır.

3.2. Derlem Veritabanının Yapısı

Derlem metinleri ve sorgulamada kullanılan evrik dizin (İng. inverted index) yapısı MySQL veritabanı yönetim sisteminde hazırlanmış bir veritabanında tutulmaktadır. Aramayı hızlandırmak ve tam metin (İng. full text) aramalarını da destekleyebilmek için MySQL veritabanı yönetim sistemindeki varsayılan veri depolama motoru olan MyISAM yapısı kullanılmıştır. Derlem veritabanı 1., 2., ve BCNF normal formlarının kurallarına uygun olarak tasarlanmıştır. Veritabanına veri ekleme, silme, güncelleme işlemleri, hazırlanan yönetici paneli aracılığıyla yapılmakta, böylece veritabanında yer alan verinin tutarlı olması da sağlanmaktadır. Derlem veritabanında bulunan tablolar ve içerdikleri veri miktarı Tablo 2'de yer almaktadır.

Tablo Adı	Kayıt Sayısı	Veri Miktarı	Açıklama
k_kitle	4	< 1 KB	Kitle Türleri
k_yazarlar	3146	< 1 KB	Yazarlar
k_yazar_turu	3	< 1 KB	Yazar Türleri
k_turev	6	< 1 KB	Türev Metin Biçimi
k_alan	9	< 1 KB	Alan
k_tur	39	< 1 KB	Tür
k_medya	4	< 1 KB	Medya
k_yayimci	672	< 1 KB	Yayıncı
k_cinsiyet	3	< 1 KB	Yazar Cinsiyeti
kunyeler_metin	4442	574 KB	Doküman Künyeleri
metinler	4442	391.3 MB	Dokümanlar
sozcukbirim	57,998,615	1.27 GB	Dizinler
teksozcuk	1,457,752	40.87 MB	Tek sözcükler

Tablo 2. TUD-Tanıtım Sürümü Veritabanı Yapısı

k_kitle tablosunda derlemde bulunan metinlerin okuyucu kitlesi türleri yer almaktadır. Derlemdeki metinlerin "çocuk", "genç", "yetişkin", "tümü" olmak üzere 4 tür okuyucusu bulunmaktadır. Böylelikle derlemde yapılacak sorgulamalarda okuyucu kitlesi türüne göre bir filtreleme yapılabilmektedir. **k_yazarlar** tablosunda ise derlemdeki tüm metinlerin yazarlarının bir listesi bulunmaktadır. **k_yazar_turu** tablosunda derlemde bulunan metinlerin yazarlarının türleri bulunmaktadır. Yazar türleri "çoklu", "kurumsal", "tekil" olabilmekte ve buna göre sorgu sonuçları filtrelenebilmektedir. **k_turev** tablosunda "bilimsel düzyazı", "kurgu ve şiir", "bilimsel olmayan düzyazı ve özyaşam", "gazete", "diğer yazılı basılmış metin", "basılmamış yazılı metin" olmak üzere türev metin biçimleri yer almakta ve buna göre sorgu sonuçlarının filtrelenmesine izin verilebilmektedir. **k_alan** tablosunda, metinlerin Tablo 1'de verilen konu alanları bulunmakta ve alana göre sorgu sonuçları filtrelenebilmektedir. **k_tür** tablosunda derlemde bulunan metinler için tanımlanmış 39 adet metin türü bulunmakta ve sorgu sonuçları bu tabloda bulunan türlere göre sınırlandırılabilmektedir. **k_medya** tablosunda derlemde bulunan belgelerin "kitap", "süreli yayın", "çeşitli:yayınlanmış", "çeşitli:yayınlanmamış" olmak üzere medya türleri bulunmakta ve bu türlere göre sorgu sonuçları filtrelenebilmektedir. **k_yayimci** tablosu derlemde bulunan metinlerin yayınevi bilgisini; **k_cinsiyet** tablosu ise derlemde bulunan metinlerin yazarlarının cinsiyet türlerini saklar. Böylece yazar cinsiyetine göre sorgu sonuçlarını filtreleme imkânı verir. **kunyeler_metin** tablosu derlemde bulunan 4442 adet metin belgesinin medya, konu alanı, yazar, yayınevi gibi künye bilgilerini saklar. **metinler** tablosunda derlemde bulunan 4442 adet belgenin tam metni yer alır. **sozcukbirim** tablosunda bölüm 3.3'de anlatılan "sözcükbirim ve teksözcük belirleme" algoritmasına göre tüm derlemden çıkarılmış sözcükbirimler ve bu sözcükbirimlerin teksözcük numarası, derlemde geçen orijinal hali, hangi belgede, hangi pozisyonda bulunduğu bilgisi yer alır. Sorgulamalarda kullanılan

ana tablolardan biridir. **teksozcuk** tablosunda da bölüm 3.3'de anlatılan "sözcükbirim ve teksözcük belirleme" algoritmasına göre tüm derlemden çıkarılmış teksözcükler, teksözcüğün numarası (birincil anahtar), türü (kelime, noktalama işareti, diğer) ve derlemdeki sayısı yer almaktadır.

3.3. Sözcükbirimleştirme (İng. tokenization) ve Evrik Dizin (İng. Inverted Index) Yapısı

Veritabanında **metinler** tablosunda 4442 adet derlem metni bulunmaktadır. Derlem üzerinde sorgulama yapabilmek için bu metinlerin içinde yer alan sözcükbirimlerin belirlenmesi, tüm derlemde yer alan teksözcüklerin çıkarılması ve bir çeşit evrik dizin yapısında hangi teksözcüğün hangi metin belgesinde ve hangi pozisyonda geçtiği bilgisinin tutulması gerekmektedir. Bu işlemleri gerçekleştirebilmek amacıyla Şekil 1'de yer alan "sözcükbirim ve teksözcük belirleme" algoritması tasarlanmış ve kullanılmıştır.

Sözcükbirim ve teksözcük belirleme algoritması FreeBSD sunucu ortamında Perl betik dili ile kodlanmıştır. Perl dili ile yazılmış sözcükbirim ve teksözcük belirleme programı MySQL veritabanına bağlanıp, metinler tablosundaki her bir metni alır, boşluklardan bölerek sözcükbirimleri oluşturur. Elde edilen sözcükbirimlerin kısaltma ya da sayısal birimler olup olmadığı kontrol edilir. Bu amaçla daha önceden belirlenmiş ve Türkçe metinlerde sıklıkla görülen kısaltmaların bir listesi kullanılmıştır. Eğer sözcükbirim bir kısaltma ya da sayısal bir ifade ise hiçbir ilave dönüşüm yapılmadan olduğu gibi alınır. Örneğin *1,000*, *13:48*, *27Temmuz2012* gibi sayısal karakter içeren ifadeler veya *P.T.T.* gibi kısaltma içeren sözcükbirimler olduğu gibi alınır.

Eğer elde edilen sözcükbirim kısaltma ya da sayısal bir ifade değilse, bu sözcükbirimin başında ya da sonunda noktalama işaretleri varsa, bu noktalama işaretleri de ayrılarak, noktalama işaretlerinin her biri ayrı bir sözcükbirim olarak alınır. Örneğin *güzellikler!* şeklindeki bir sözcükbirim *güzellikler* ve *!* şeklinde 2 adet sözcükbirime ayrılır. Elde edilen sözcükbirim, kısaltma ve sayısal karakter içeren hariç, küçük harfe dönüştürülür. Bu dönüşümden sonra oluşan sözcükbirim önce *teksözcük* tablosundan aranır. Eğer *teksözcük* tablosunda varsa, bu sözcükbirim derlemde daha önce elde edilmiş demektir. Bu durumda bu teksözcüğe atanmış *teksözcük_no* değeri alınır, bu teksözcüğün sayaç değeri 1 attırılır, *sözcükbirim* tablosuna ise bulunan bu sözcükbirim metinde geçen haliyle (küçük harf dönüşümü yapılmadan) eklenir. Ayrıca elde edilen sözcükbirimin *teksözcük_no* değeri, hangi belgede hangi pozisyonda geçtiği bilgileri de *sözcükbirim* tablosuna eklenir. Eğer oluşturulan sözcükbirim *teksözcük* tablosunda yoksa, önce *teksözcük* tablosuna eklenir. Bu teksözcük için bir *teksözcük_no* değeri verilir, *sayaç* değeri 1 yapılır ve *türü* de belirlenerek *teksözcük* tablosuna bu veriler eklenir. Daha sonra bu sözcükbirim *sözcükbirim* tablosuna hangi belgede, hangi pozisyonda geçtiği bilgisiyle eklenir.

Kısaltma ve sayısal ifadelerin dışında kalan sözcükbirimler küçük harfe çevrilerek *teksözcük* tablosuna eklenmiş, ancak *sözcükbirim* tablosuna ise metinde geçtiği orijinal haliyle eklenmiştir. Böylece, sorgulama sırasında büyük/küçük harf ayrımı yapmadan ya da yaparak her iki şekilde de arama yapmak mümkün olabilmektedir. Ancak sözcükbirim içinde geçen noktalama işaretleri ayrılmamıştır. Örneğin *siyah-beyaz* veya *Adana'nın* sözcükbirimleri sadece küçük harfe dönüştürme yaparak olduğu gibi teksözcük olarak alınmıştır.

Sözcükbirim ve teksözcük belirleme algoritmasına göre 4442 doküman bulunan derlemden 57,998,615 adet sözcükbirim elde edilmiş olup, bu sözcükbirimlerin yaklaşık 48 milyon adedi noktalama işareti haricindeki sözcükbirimlerdir. TUD-Tanıtım Sürümü için toplam 1,457,752 adet teksözcük belirlenmiştir.

Sözcükbirim ve teksözcük belirleme programının hızlı çalışması için *sözcükbirim* ve *teksöz-*

cük tabloları bellekte çırpı tablosu (İng. hash table) olarak tutulmuş olup, daha sonra işlemler bittikten sonra CSV uzantılı olarak sabit diske kaydedilmiştir. Bu işlemler 4442 doküman için sunucu ortamında ortalama 1189,2 saniye sürmüştür. Elde edilen CSV dosyaları veritabanda *teksözcük* ve *sözcükbirim* isimli tablolara aktarılarak işlemler tamamlanmıştır.

Algoritma: Sözcükbirim ve Teksözcük belirleme

Input: *metinler* tablosu, *kısaltmalar listesi, noktalama işaretleri listesi*
Output: *sözcükbirim* ve *teksözcük* tabloları

metinler tablosundaki her *metin* için:
1. *metin* boşluklardan bölünerek sözcükbirimler elde edilir ve bir sözcükbirim (*S*) listesine eklenir.
2. *S* listesindeki her sözcükbirim (*s*) için
   - Eğer (*s* ∈ *kısaltmalar listesi*) || ([0-1] ⊂ *s*) ise,
     i) *s*'yi *teksözcük* tablosunda ara, eğer varsa *teksözcük_no*'yu al, yoksa <*s, teksözcük_no, sayaç, tür*> kaydını *teksözcuük* tablosuna ekle.
     ii) <*s, teksözcük_no, metin_no, pozisyon_no*> kaydını *sözcükbirim* tablosuna ekle.
   - Eğer *s*'nin başında ya da sonunda noktalama işareti varsa, *s*' yi küçük harflere çevir, tüm noktalama işaretlerini ayır, elde edilen her sözcükbirim için *i* ve *ii* adımlarındaki işlemleri tekrarla.
   - Eğer *s*'nin başında ya da sonunda noktalama işareti yoksa, *s*' yi küçük harflere çevir, ve elde edilen sözcük birim için *i* ve *ii* adımlarındaki işlemleri tekrarla.

Şekil 1. Sözcükbirim ve Teksözcük Belirleme Algoritması

3.4 Sorgulama ve Ön Belleğe Yükleme

Derlemin web arayüzü sunucuda çalışmaya başladığı anda veritabanında yer alan *teksözcük* tablosu RAM-Belleğe aktarılmaktadır. Bu aktarım APC uzantısı [18] ile PHP ara yüzünden yapılmaktadır. APC (Alternative PHP Cache), veri tabanında veya dosyalarda bulunan ve uygulama sırasında sıkça erişilen bilgileri, RAM'da saklama yöntemi ile bir nevi hız ve optimizasyon sağlama aracıdır.

Derlem sorgularının kullanıcıya daha hızlı bir şekilde ulaştırılması için *teksözcük* tablosunda yer alan teksözcüklerin *sözcükbirim* tablosunda yer alan sorgu sonuçları önceden hesaplanmış ve sabit diskte metin belgelerinde saklanmıştır. Bu sonuç dosyalarında bir sorgu terimine ait "bağımlı dizin" dizilimi ve sonuçları yer almaktadır. Bu dosyaların diskte kapladığı alanın azaltılması ve diskten okunması işlemleri için "igbinary" serialize [19] yöntemi uygulanmıştır. Yer kazanımı ve bellek kullanımında etkili sonuçlar vermiştir. Derlemin yeni sürümünde sorgu sonuçlarının *sözcükbirim* tablosu üzerinden gerçek zamanlı hesaplanması planlanmaktadır.

Kullanıcı tarafından girilen bir teksözcük belleğe önceden aktarılmış *teksözcük* tablosundan hızlı bir şekilde aranır ve o sorguya ait *teksözcük_no* değeri bulunup, o teksözcüğe ait daha önce hesaplanmış sonuç dosyası diskten alınarak işlenmek ve ekranda görüntülenmek üzere arayüze gönderilir. Kullanıcının belirlediği filtreleme ölçütlerine göre, sonuç dizilimi belirlenir ve bu dizilim rastgele sıralanır, daha sonra yazılım tarafından ön tanımlı olarak 2.500 sonuç ekranda görüntülenir.

Görüntüleme sırasında SpryMedia [9] tarafından geliştirilmiş DataTables kullanılmaktadır. Bu sayede sonuçlar görsel açıdan hızlı ve etkin biçimde görüntülenebilmektedir. Bu işlemlerin yanı sıra kullanıcının daha sonra talep edebileceği sıralama ve eşdizimlilik işlemleri için de elde edilen sonuçlar önbelleğe (İng. cache) alınmaktadır.

APC uzantısı bilgileri bellekte az yer kaplaması için serialize eder. Bu serialize işlemleri için yine "igbinary" [19] serializing yöntemi kullanılmıştır. Bu sayede normal serialize yöntemlerine göre hız ve boyut açısından kazanım olmuştur. PHP, igbinary yöntemi ile normal saklama ve serialize işlemine göre yaklaşık 1/5 oranında yer kazanımı sağlamıştır [12].

4. TUD- Tanıtım Sürümü Arayüz Özellikleri

TUD-Tanıtım Sürümü temel olarak tek sözcük ya da sözcük grubunun bağlam içinde anahtar sözcük (İng. KWIC) arama işlevine sahiptir. Bununla birlikte, araştırmacılar derlem anasay-

fasında bulunan çeşitli dil dışı ölçütlerle (yayın yılı, alan, türev metin biçimi, vb.) araştırma sorularına uygun olarak sorgularını daraltabilir ve bu doğrultuda bağımlı dizin sonuçlarına, listeleme işlevine ve sayısal sıralı eşdizimlilik listelerine ulaşabilirler. 3. bölümde yazılım mimarisi anlatılan TUD- Tanıtım Sürümü arayüzünün sahip olduğu işlevler aşağıda sıralanmıştır.

1. "Yayın yılı, medya, metin örneklemi, alan, türev metin biçimi, yazarın cinsiyeti, yazar ya da yazarların türü, okuyucu kitlesi ve tür" ölçütlerine bağlı olarak aramalarını daraltabilir ve belirledikleri ölçütlere göre bağımlı dizin sonuçları alabilirler.
2. Arayüzün listelediği bağımlı dizin sorgularındaki sorgu sözcüğü ya da sözcüklerinin $\pm$ 35 sözcüklük bağlamına erişebilirler.
3. Listelenen bağımlı dizinlerin geçtiği metinlerin veribilgisine erişebilirler.
4. Sorgu sonuçlarını Excel ve metin dosyası formatında dışa aktarabilirler.
5. Sorgu teriminin "Türev metin biçimi, alan, okuyucu kitlesi, medya, cinsiyet ve yıl" ölçütlerine göre "sözcük sayısını, eşleşme sayısını, metinlerdeki dağılımını," bir milyon sözcükteki sıklığını görüntüleyebilirler.
6. Sorgu sözcüğünün solundaki ve sağındaki $\pm$ 5 sözcüğe göre alfabetik listelerini alabilirler.
7. Sorgu sözcüğünün çoğunlukla hangi sözcüklerle ya da dilbilgisi ulamlarıyla birlikte olduğunu, çeşitli istatiksel hesaplamalar kullanarak (LL, MI, MI3, T, Dice coefficient, Logdice coefficient değerleri) $\pm$ 5 sözcüklük aralıkta düzenlenmiş, sayısal sıralı eşdizimlilik listeleri şeklinde alabilirler.

5. Sonuç

Bu çalışmada güncel Türkçenin web tabanlı ilk referans derlemi Türkçe Ulusal Derlemi'nin tasarım ölçütleri, yazılım mimarisi, derlem veritabanı yapısı, sözcükbirimleştirme ve evrik dizin yapısıyla derlem verisinin işlenmesi ve TUD-Tanıtım Sürümü'nün kullanıcılara sunduğu arayüz özellikleri gösterilmiştir.

TUD bilişim, eğitim, medya, Türkçenin güncel kullanımı ve tanıtımı ile ilgili tüm kişi ve kurumların kolayca ulaşabileceği ve yararlanabileceği bir dil kaynağıdır. Bir sözcüğün, kalıp sözün, bileşik yapının ya da dilbilimsel bir birimin/ekin kullanım sıklığının ve farklı ortamlardaki görünüm ve işlevlerinin bulunması; bilgisayarlı çeviriden sözlük hazırlamaya, Türkçenin anadil ya da yabancı dil olarak öğretimi için araçlar geliştirmeye, elektronik şifreleme ve arşiv oluşturmaya, dilin sözvarlığındaki değişmeyi ve çeşitliliği saptamaya kadar uzanacak geniş bir alanda uygulamalar için veri sağlayacak niteliktedir.

Dilbilimcilerin ve bilgisayar mühendislerinin disiplinler arası çalışmasıyla TUD veritabanları kullanılarak, TÜBİTAK (Proje no:113K039) destekli yeni bir proje sürmektedir. Proje 49 milyon sözcüklük yazılı Türkçe metin temelinde, TUD'da bulunan tüm sözcüklerin sözcük türlerini ve ek özelliklerini gösteren, Türkçe için bir ilk olacak, güncel bir doğal dil işleme (DDİ) sözlüğü oluşturmayı ve oluşturulan sözlük yardımıyla otomatik işaretlenen TUD metinleri kullanılarak derlem-temelli bir sözcük ve ek sıklığı sözlüğü hazırlamayı hedeflemektedir. Buna paralel olarak, oluşturulacak DDİ sözlüğünü yazılımında kullanan, herkesin erişebileceği, web tabanlı bir morfolojik işaretleyici tasarlanacak, geliştirilecek ve çevrimiçi, ücretsiz biçimde son kullanıcının hizmetine sunulacaktır.

6. Kaynaklar

[1] Aksan, Y. et al., "Construction of the Turkish National Corpus (TNC)", **Proceeding of the Eight International Conference on Language Resources and Evaluation (LREC 2012),** İstanbul, (2012).

[2] Ambati, B., Reddy, S., Kilgarriff, A., "Word sketches for Turkish", **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)** İstanbul, (2012).

[3] Apache HTTP Server Project, http://httpd.apache.org/

[4] Aston, G., Burnard, L., "The BNC handbook: Exploring the British National Corpus with SARA", Edinburgh: Edinburgh University Press. (1998).

[5] Baisa, V. ve Suchomel., V., "Large corpora for Turkic Languages and unsupervised morphological analysis", **Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)**, İstanbul, (2012).

[6] Biber, D., Representativeness in corpus design. Literary and Linguistic Computing, 8, 243-257 (1993).

[7] CSS, http://www.w3schools.com/css/

[8] Dalkılıç, G., Çebi,Y., A 300 mb turkish corpus and word analysis, Advances in information system, 205–212, (2002).

[9] Datatables by SpryMedia, http://www.sprymedia.co.uk/article/DataTables

[10] HTML, http://www.w3schools.com/html/

[11] Hundt, M., Nesselhauf, N. ve Biewer, C. (Eds.), "Corpus linguistics and the web", Amsterdam/New York: Rodopi (2007).

[12] Javascript,http://www.w3schools.com/js/

[13] Jquery, http://jquery.com/

[14] Lee, D., "What corpora are available?" A. O'Keefe ve M. McCarthy, (Eds.), The Routledge handbook of corpus linguistics, 107-121, London: Routldege, (2012).

[15] McEnery, T., Hardie, A., "Corpus linguistics", Cambridge: Cambridge University Press, (2012).

[16] MySQL 5.5 Release Notes, http://dev.mysql.com/doc/relnotes/mysql/5.5/en/

[17] PHP5.4.2, http://www.php.net/releases/5_4_21.php

[18] PHP APC Extension, http://php.net/manual/en/book.apc.php

[19] PHP PECL IGBinary Extension, http://codepoets.co.uk/2011/php-serialization-igbinary/

[20] Sak, H., Güngör, T., Saraçlar, M., "Turkish language resources: Morphological parser, morphological disambiguator and web corpus", **Advances in natural language processing**, 417–427, (2008).

[21] Say, B., Zeyrek, D., Oflazer, K. ve Özge, U., "Development of a corpus and a treebank for present-day written Turkish", Current research in Turkish linguistics: proceedings of the 11th International Conference of Turkish Linguistics, 183-192, (2002).

[22] Sinclair, J. M. ,"How to build a corpus", M. Wynee (Ed.), Developing linguistic corpora: A guide to good practice, ss. 96-101, (2005).

[23] The FreeBSD Project, http://www.freebsd.org/

[24] The Perl Programming Language, http://www.perl.org/get.html

[25] Wynne, J. (Ed.), "Developing linguistic corpora: A guide to good practice", http://www.ahds.ac.uk/guides/linguistic-corpora/appendix , (2005).