KMÜ Sosyal ve Ekonomik Araştırmalar Dergisi 16 (Özel Sayı I): 181-188, 2014
ISSN: 2147 - 7833, www.kmu.edu.tr

Türkçe Çevirimiçi Derlemler Üzerine

Serdar Karaoğlu
Afyon Kocatepe Üniversitesi, Fen-Edebiyat Fakültesi, Afyonkarahisar

Özet

Teknolojik gelişmeler her alanda olduğu gibi dil çalışmalarındaki araştırma yöntemlerinde de bir takım değişiklikler meydana getirmektedir. Bu değişiklikler doğrultusunda, Bilgisayarlı Dilbilim gibi disiplinlerarası çalışma alanları oluşmakta ve araştırmacılara büyük kolaylıklar sunan sonuçlar elde edilmektedir. Söz gelimi, bu olanaklardan birisi olduğunu düşündüğümüz çevirimiçi derlemler, bir araştırmacının herhangi bir biçimbirimin işlevlerini tespit edebilmek için aylarca süren el ile fişleme yöntemine nazaran daha önce elektronik ortama aktarılmış ve işlenilmiş olan veri tabanı sayesinde birkaç saniye içinde aynı neticeyi verebilmektedir. Biz de bu bağlamda, dünya ve Türkiye'deki çevirimiçi derlem çalışmalarından kısaca bahsederek Türkçe çevirimiçi derlemler hakkında bilgi vereceğiz.

Anahtar Kelimeler: Bilgisayarlı dilbilim, Türkçe çevirimiçi derlem, biçimbirim.

The Relationship Between The Structure And Socio-Culturel Crime Case: The Case Of Malatya

Abstract

Tecnologichal developments occur several kinds of changes in research methods of linguistic area as in all areas. Through the changes, some interdisciplinary fields like computational linguistics are emerging and some results those provides great conveniences to researchers are obtained. For instance, on the contrary of preparing an index card by hand to identify the functions of any gramatical morpheme during months, Turkish online corpus which is considered to be one of these conveniences gives the same result in a few seconds thanks to its processed database in the electronic environment. In this context, we will give an information on the world and Turkish online corpuses by shortly mentioning on the approaches in Turkey about online corpuses.

Keywords: Computational linguistics, Turkish online corpus, gramatical morpheme.

1. Giriş

Teknolojik gelişmeler her alanda olduğu gibi dil çalışmalarındaki araştırma yöntemlerinde de bir takım değişiklikler meydana getirmektedir. Bu değişiklikler doğrultusunda, Bilgisayarlı Dilbilim gibi disiplinlerarası çalışma alanları oluşmakta ve araştırmacılara büyük kolaylıklar sunan sonuçlar elde edilmektedir.

Dilin canlı bir varlık olduğu ve bunun neticesinde sürekli geliştiği/değiştiği tüm araştırmacılar tarafından kabul edilen bir olgudur. Sürekli gelişen/değişen bu canlı varlık, derlem vasıtasıyla değerlendirilebilmekte, sözlükte yer alacak madde başları tespit edilebilmekte ve bu madde başlarının anlam çeşitliliği saptana bilmekte, dilbilgisi araştırmalarında herhangi bir biçimbirimin işlevleri daha kolay tespit edilebilmekte ve edimbilim, anlambilim gibi dilbilim alanları ile ilgili araştırmalar yapılabilmektedir (McEnery vd., 2001: 103-132).

Belli başlı faydalarından bir kaçını sıraladığımız derlem, "belirli bir dilin yazılı metinlerinin ya da konuşma dilinin bir takım kıstaslar çerçevesinde o dili temsil edebilecek dile ait verilerin bir araya getirilmesi" olarak tanımlanmaktadır (Burkhanov, 49-50). Derlemler elektronik ve elektronik olmayan derlemler şeklinde iki gruba ayrılabilir. Elektronik olmayan derlemler genel itibariyle bireylerin kendi araştırmalarında kullanmak için dile ait verileri bir araya getirmesiyle oluşan derlemlerdir. Bu tür derlemler çoğunlukla derlemi oluşturan kişi(ler)de saklı kaldığı için diğer araştırmacılar tarafından kullanımı oldukça zor olmaktadır. Ancak bunun yanında, oluşturulan çalışmaların yayını sayesinde araştırmacıların kolaylıkla erişebileceği derlemler de bulunmaktadır[1] . Elektronik bir veritabanında bir dilin yazılı metin ya da sözlü dil parçalarının belirli kurallar çerçevesinde tümünün ya da bir kısmının bir araya getirilmesiyle oluşturulan derlemlere ise elektronik derlem denilmektedir (Kennedy, 1998).

Elektronik derlemler sanal ortamda erişime açık olup olmamalarına göre "çevirimiçi (online)" ve "çevirimdışı (offline)" olarak ikiye ayrılmaktadır. Çevirimdışı derlemler Cd-ROM ya da kasete kaydedilmiş verilerin bilgisayara/teybe kurulması/takılmasıyla çalışan derlemler; çevirimiçi derlemler ise sanal ortamda kullanıma açık olan derlemlerdir (McEnery, vd., 2012: 6-7; Kennedy, 1998: 13).

Bu çalışmada Türkçe çevirimiçi derlemlerden yazılı metinlerin bir araya getirilmesiyle oluşturulan derlemler incelenecektir. Türkçe çevirimiçi derlemlere geçmeden önce dünya dilleri üzerine yapılmış derlemlere kısaca değinilecektir.

[1] 1980'lerde Dimitriy Vasilyev'in hazırladığı Yenisey Yazıtlarını konu alan "Korpus Turkskih Runiçeskih Pamyatnikov Basseyna Yeniseya", Leningrad 1983 ve Karaçay bilgini S. Ya. Bayçorov tarafından yayınlanan Avrupa'daki Köktürk harfli metinlerin derlemi sayıla bilecek "Drevnetyurkskiye Runiçeskiye Pamyatniki Yevropı" adlı eser bu tür derlemlere örnek verilebilir. Bayçorov'un eseri Muvaffak Dumanlı tarafından "Avrupa'nın Eski Türk Runik Abideleri" adıyla Türkçeye çevrilmiştir. Daha geniş bilgi için bkz.

2. Dünya Dilleri Üzerine Yapılmış Derlem Çalışmaları

Derlem dilbilimi (corpus linguistics) Noam Chomsky'nin modern dil çalışmalarında çığır açan 1957 yılındaki "Sözdizimi yapıları (Syntactic Structures)" adlı çalışmasından sonra 1960'larda ortaya çıkmıştır. Chomsky bu çalışmasıyla dilbilim literatüründe kabul görmüş teorilerin yeniden gözden geçirilmesi zorunluluğunu ortaya koymuştur. Giderek evrensel bir fenomen olan bu teoriye bağlı olarak araştırmacılar ilgilendikleri dillerle alakalı daha fazla tatmin edici sonuç almaya başlamalarına rağmen dille ilgili yeterince metne ulaşamadıkları için yaptıkları bazı tanım ve kurallar geçersiz olabilmiştir. 1950'lerde yalnızca içgözlemle cevaplanılamayacak deneye dayalı (betimsel) soruların ortaya çıkması dilin gerçek verilerine olan ihtiyacı ortaya koymuştur.

Betimsel dil araştırması için dil verilerini bir araya getiren geniş çaplı ilk proje Randolph Quirk's Survey of English-speaking Usage'dir. Bu çalışma bilgisayar ortamında oluşturulmuş bir çalışma değildir. Betimsel dil çalışmaları için temel bir yapı oluşturan bu çalışma sözlü ve yazılı verilerden oluşan bir milyon sözcük içeren bir çalışmadır.

1960'larda ikinci veritabanlı proje Brown Corpus'tur. On beş farklı türe ait beş yüz Amerikan metinlerinden (kitaplarından) iki bin sözcüklü örneğin bir araya getirildiği bir milyon sözcüklü derlemdir. Nelson Francis tarafından elektronik verilere derlem (corpus) teriminin ilk uygulandığı çalışmadır. Elle işaretlenen/işlenen bu derlem hem dilbilgisi hem de sözlük ile ilgili sorulara cevap vereceği düşünülürken tüm söz varlığının küçük bir kesiti olan böyle bir derlemin aranan cevapları karşılamayacağı daha sonra anlaşılmıştır.

1970'lerde British English (Brown Corpus)'e benzer tarzda oluşturulmuş olan LOB (Lancester-Oslo-Bergen) Corpus sözcük sıklığı ve dilbilgisi çalışmaları için önemli bir kaynak olarak görülmüştür. Ancak Brown Corpus gibi bu çalışmada dilin küçük bir kesitini yansıttığı için aranan sorulara tam olarak cevap verecek kapasitede olmamıştır (M. A. K Halliday, vd., 2004: 107-112).

1980 itibariyle bilgisayar destekli derlem çalışmaları hız kazanmıştır. Örnek olarak verebileceğimiz belli başlı çalışmalar şunlardır:

-British National Corpus (BNC) http://info.ox.ac.uk/bnc

-Contemporarary Corpus of American English (COCA)

-International Corpus of English (ICE) http://www.ucl.ac.uk/english-usage/ice/

-American National Corpus (ANC) http://www.americannationalcorpus.org/OANC/index.html

-Polish National Corpus http://nkjp.pl/index.php?page=0&lang=1

-Russian Reference Corpus http://www.ruscorpora.ru/en/

-Corpus Del Español http://www.corpusdelespanol.org

-A Historical Corpus of the Welsh Language 1500-1850 http://people.ds.cam.ac.uk/dwew2/hcwl/menu.htm

3. Türkçe Üzerine Yapılmış Derlem Çalışmaları

Türk Dili üzerine derlem çalışmalarından ilki Bilge Say başkanlığında gerçekleştirilen ODTÜ Türkçe Derlem olarak da bilinen "Bilgisayar Ortamında Bir Derlem Geliştirme Çalışması"dır. Derlemde 1990 sonrası sadece yazılı dili içeren metinler bulunmaktadır. Sözlü dile ait veri yoktur. Farklı türde metinlerden 2000 sözcüklü örneklemlerin seçilip elektronik ortama aktarılarak işaretlenmesi ile oluşturulan iki milyon sözcüklü çevirimdışı bir derlem (treebank)'dir (Say vd, 2002: 183-192).

Türkiye Türkçesi üzerine yapılmış bir diğer derlem çalışması "Türkçe Ulusal Derlem"dir. TÜBİTAK tarafından desteklenen bu proje Mersin Üniversitesi Dilbilim alanı araştırmacıları tarafından dizayn edilmiştir. 2008 yılında başlanan bu çalışmanın tanıtım sürümü 2012'de kullanıma sunulmuştur. 50 milyon sözcük kapasiteli, 1990-2009 yılları arasında farklı alan ve türlerde %95'i yazılı, %5'i sözlü örnekleri içeren dengeli, karışık (yazılı-sözlü), eşzamanlı ve genel bir derlemdir (Aksan vd., 2009: 300).

Taner Sezer tarafından hazırlanan 491 milyon sözcükten oluşan genel TS Corpus'un ilk sürümü 1 Mart 2012, ikinci sürümü 30 Ağustos 2012'de çevirimiçi olarak kullanıma sunulmuştur. TS Corpus sözcük türü, biçimbirim ve kök sözcük etiketlemesi yapılarak hazırlanan ve kullanıcıya büyük kolaylıklar sunan genel amaçlı bir derlemdir[2] .

Türkçenin artsüremli/tarihsel derlemi olan Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D) Orhon Türkçesi, Uygur Türkçesi ve Karahanlı Türkçesi'ne ait yazılı metinlerin elektronik ortama aktarılarak sözcükbirim ve sözdizim bazında işaretlenmesiyle oluşturulmuş 600 yıllık bir dönemi (7-13. yy) kapsayan 400-450 bin sözcük içeren çevirimiçi derlemdir[3].

Son olarak vereceğimiz çevirimiçi yazılı metin derlem "Vorislamische Alttürkische Texte: Elektronisches Corpus 'VATEC' (İslamiyet Öncesi Türkçe Metinleri Elektronik Derlem)"i 1999-2003 yılları arasında Deutsche Forschungsgemeinschaft tarafından desteklenen ve Prof. Dr. Marcel Erdal başkanlığında gerçekleştirilen bir projedir[4] . Derlem web sayfası Almanca olarak sunulmuştur. Derlem içerisinde Uygur Türkçesi dönemine ait metinler yer almaktadır.

2008-2010 yılları arasında TÜBİTAK tarafından desteklenen "Sözlü Türkçe Derlemi (STD)" yüz yüze ya da çeşitli iletişim araçlarıyla (örn. telefon ve kitle iletişim araçları) gerçekleşen Türkçe konuşmalardan oluşan 1 milyon sözcüklük veritabanının dilbilimsel yöntemlerle çözümlenerek günümüz Türkçesinin bilgisayar ortamında izlenebilmesini amaçlayan çevirimiçi bir derlemdir. 2010 yılında derlemin deneme sürümü kullanıcılara araştırmaları için sunulmuş olup 2013'ün sonuna doğru ise konuşmaların yazıya çevrilmiş hali ile 400.000 sözcüklük sürümünün kullanıma sunulması düşünülmektedir[5].

---

[2] http://www.tscorpus.com/tr (çevirimiçi) 30.03.2013

[3] http://derlem.cu.edu.tr/index.php?a=tarihsel/icerik_amac (çevirimiçi) 30.03.2013

[4] http://vatec2.fkidg1.uni-frankfurt.de/ (çevirimiçi) 30.03.2013

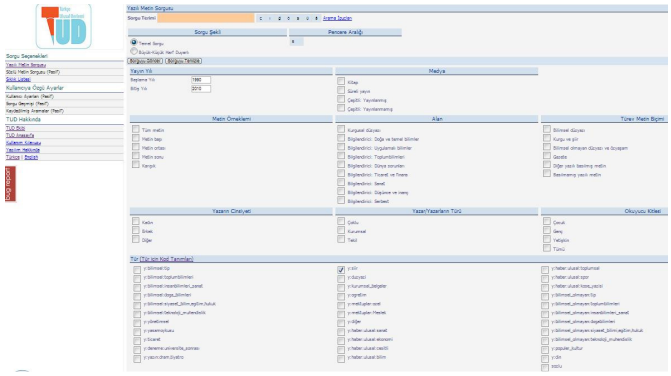
[5] http://stc.org.tr/hakkinda/ (çevirimiçi) 30.03.2013

3.1. Türkçe Çevirimiçi Derlemlerin Araştırmacılara Sunduğu (Sorgu) Olanaklar(ı)

Bildirimizin bu bölümünde çevirimiçi olarak kullanıma sunulan "Türkçe Ulusal Derlem (TUD), TS Corpus ve Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D) ve Vorislamische Alttürkische Texte: Elektronisches Corpus (VATEC)"un kullanıcılara ne tür sorgu olanakları sundukları irdelenecektir[6].

Mersin Üniversitesi Dilbilim Araştırmacılarınca (Yeşim Aksan vd.) 2008'de başlatılan Türkçe Ulusal Derlem (TUD) projesinin 2012 yılında tanıtım sürümü çevirimiçi olarak kullanıma sunulmuştur. http://www.tnc.org.tr/ adresinden ulaşılabilinen derleme kayıt olunarak sağlayıcı tarafından kullanıcı adına oluşturulan şifre vasıtasıyla derlemin tanıtım sürümüne erişilebilinmektedir. Kullanıcı adına oluşturulan şifre ile tanıtım sürümünü açtığımızda karşımıza Tablo-I'deki Tanıtım Sürümü ana sayfası gelmektedir.

Tablo-1: Tanıtım Sürümü Ana Sayfası



Tanıtım sürümü ana sayfasının sol tarafında "Sorgu Seçenekleri", "Kullanıcıya Özgü Ayarlar" ve "TUD Hakkında" menü seçenekleri bulunmaktadır.

Sorgu seçenekleri menüsünde %95'i yazılı metin %5'i sözlü metin olarak tasarlanan TUD'in sözlü metin kısmı pasif konumdadır. Kullanıcılar şu durumda sadece yazılı metinlerden oluşan 'Tanıtım Sürümü'nde sorgulama yapabilmektedirler. Bu bölümde bulunan "Sıklık Listesi" butonu tıklandığında pdf. formatında "TUD Tanıtım Sürümü Sözcük Sıklığı Listeleri"ne ulaşılmaktadır.

"Kullanıcıya Özgü Ayarlar" menüsünde bulunan "Kullanıcı Ayarları, Sorgu Geçmişi, Kaydedilmiş Aramalar" pasif durumdadır.

"TUD Hakkında" menüsünde ise "TUD Ekibi", "TUD Anasayfa, Kullanım Kılavuzu[7] ", "Yazılım Hakkında" ve tanıtım sürümü anasayfasının "İngilizce veya Türkçe" sunulmasını sağlayan buton bulunmaktadır.

"Yazılı Metin Sorgusu" bölümünde bulunan "Sorgu Terimi" arayüzüne sorgulatmak istediğimiz sözcük (yüz) ya da iki sözcükten oluşan sözcük öbeğini (yüz göz) yazıp "Temel Sorgu" ya da "Büyük-Küçük Harf Duyarlı" sorgu şeklini seçerek +/- 10 "Pencere Aralığı "nda[8] sorguyu gerçekleştirebiliriz.

Sorgu arayüzünün sağ tarafında bulunan "ç, ı, ğ, ö, ş, ü, â" harf listesi Türkçe klavye kullanmayan kullanıcılar için tasarlanmıştır. Bu listenin sağ tarafındaki "Arama İpuçları" butonu ise Tablo-2'de gösterildiği gibi sorgulamada faydalanılabilinecek olan joker karakterler hakkında bilgi vermektedir[9].

Tablo-2

Arama İpuçları

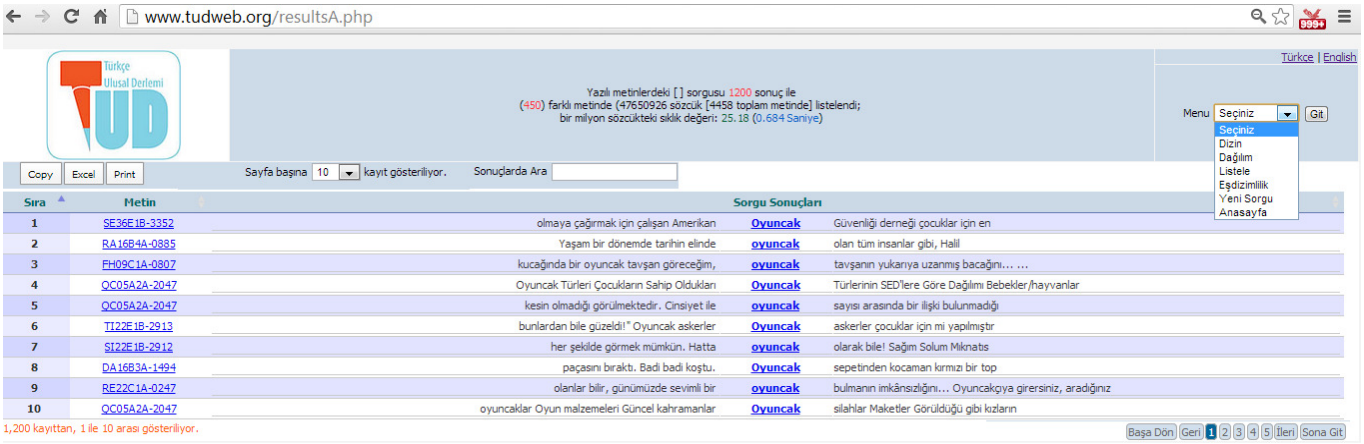
(kol) : tam arama
(kol*) : kol ile başlayanlar
(*kol) : kol ile bitenler
(*kol*) : içinde kol geçenler
(kol?uk) (a??lık) : ? yerine herhangi bir karakter gelebilir; koltuk, kolluk, aralık, aşılık vs.
(beyaz peynir) : beyaz veya peynir geçen dizilimler
("beyaz peynir") : tam olarak "beyaz peynir" dizilimi

Kullanıcılar, 1990-2010 yılları arası "kitap, süreli yayın, çeşitli yayınlanmış metinler, çeşitli yayınlanmamış metinler" türlerinden; "kurgusal düzyazı, toplum bilimleri, sanat, ticaret ve finans, düşünce ve inanç, dünya sorunları, uygulamalı bilimler, doğa ve temel bilimler ve diğer" olmak üzere dokuz farklı alanda; "yazarın cinsiyeti (kadın, erkek)", "yazar/yazarların türü (çoklu, kurumsal, tekil)" ve "okuyucu kitlesi (çocuk, genç, yetişkin, tümü)" künye bilgilerini içeren seçeneklerle sorgularını istedikleri özelliklere göre sınırlaya/genişletebilirler.

Yukarıda belirttiğimiz arama özelliklerine bağlı olarak sorgu arayüzünde "oyuncak" sözcüğünü yazarak yapacağımız sorgu sonucu Tablo-3'teki gibi ekrana yansımaktadır.

[6] Sorgu şekilleri derlemlerin web sayfalarındaki açıklamalardan faydalanılarak verilmiştir.
[7] Bu çalışmada TUD'un kullanımı ile ilgili bilgi büyük ölçüde bu kılavuzdan sağlanmıştır.
[8] Sorgu sonuç arayüzünde anahtar sözcüğün sağ ve solunda listelenecek sözcük sayısını ifade etmektedir.
[9] http://www.tudweb.org/index.php?dil=1 02.04.2013

Tablo-3



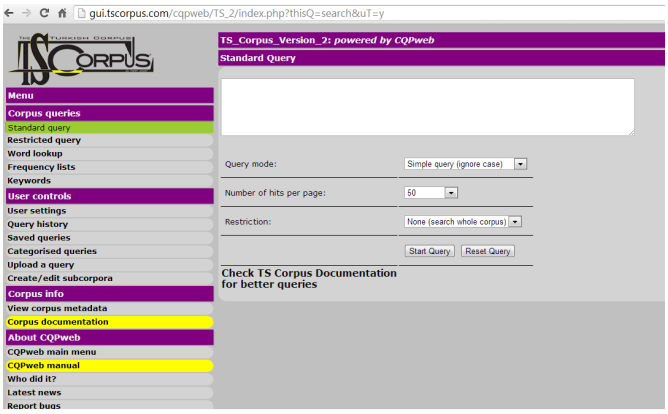
Yansıyan bu ekranın üst kısmında toplam 4458 metinde aranan sözcüğün (oyuncak), kaç farklı metinde (450), kaç kere kullanıldığı (1200) ve bu kullanımın bir milyon sözcükteki sıklık değeri verilmektedir.

Sonuçların dizin olarak verildiği bu bölümde mavi olarak sıralanan aranan sözcüğün sağ ve solunda bulunacak sözcük sayısı +/- 10 düzeyinde sorguya başlamadan önce "Pencere Aralığı" seçeneğinden ayarlanabilmektedir. Mavi renkle gösterilen "sorgulanan sözcük" tıklandığında sözcüğün geçtiği bağlam verilmektedir. Metin sütununda bulunan öğeler tıklandığında ise sözcüğün geçtiği metin hakkında bilgilere ulaşılmaktadır. Sorguladığımız sözcükle ilgili sonuçları yazıcı vasıtasıyla derlemden çekebileceğimiz gibi Exel formatında bilgisayarımıza da kayıt edebiliriz.

Ekranın sağ tarafında bulunan "menü" butonundan "dağılım, listele ve eşdizimlilik" seçenekleri vasıtasıyla sözcüğün "yayın yılı, medya, metin örneklemi, alan, türev metin biçimi, yazarın cinsiyeti, yazar/yazarların türü, okuyucu kitlesi ve tür" yönünden dağılımını; "listele" seçeneği ile anahtar sözcüğün (oyuncak) sağ ve sol tarafında bulunan sözcüklerin –istenildiğinde alfabetik sıraya göre- birer sütun oluşturularak dizilimini ve son olarak da "eşdizimliği" ile anahtar sözcüğün sağ ve/veya sol tarafında bulunan sözcüklerle kullanım sıklığını görebilmekteyiz.

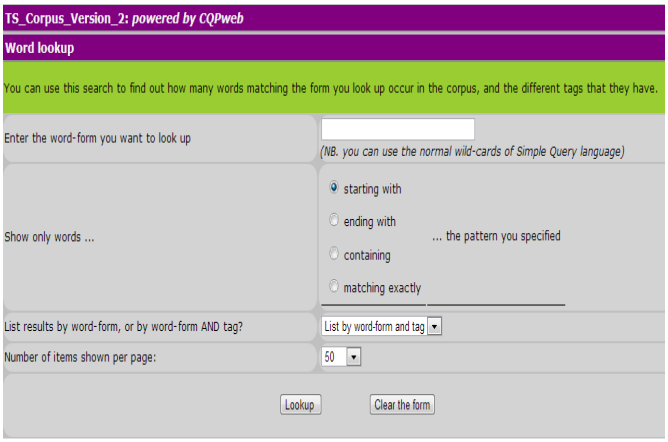
Türkçe çevirimiçi derlemlerden bir diğeri TS Corpus, Taner Sezer tarafından 491 milyon sözcük kapasiteli olarak 1 Mart 2012'de birinci sürümü, 30 Ağustos 2012'de de ikinci sürümü kullanıma sunulmuştur. TS Corpus sözcük türü, biçimbirim ve kök sözcük etiketlemesi yapılarak kullanıcıya büyük kolaylıklar sunan genel amaçlı dengesiz bir derlemdir. Derleme http://www.tscorpus.com/tr adresinden ulaşıp kayıt menüsünden anında kullanıcı adı ve şifre oluşturup Tablo-4'te verilen derlem sayfasına erişmek mümkündür.

Tablo-4



İngilizce olarak tasarlanmış olan derlem ana sayfasının sol kısmında "standart arama (standard query), sınırlandırılmış arama (restricted query), kelime arama (word look up), sıklık listeleri (frequency lists) ve anahtar sözcükler" seçeneklerini içeren "Derlem Sorgu Seçenekleri (Corpus queries)" bulunmaktadır. Bu sorgu seçeneklerinden "standart arama" Tablo-4'te verilen arama türüdür. Sınırlandırılmış arama seçeneği ile yukarıda görmüş olduğumuz sayfaya "genel veri" ya da "gazete verisi" seçenekleri gelmektedir. Toplamda iki farklı dosyadan oluşan Ts Corpus'u bu seçenekler sayesinde sınırlandırabilmekteyiz. "Kelime arama" seçeneğini tıkladığımızda ise ekrana Tablo-5'teki sayfa gelmektedir.

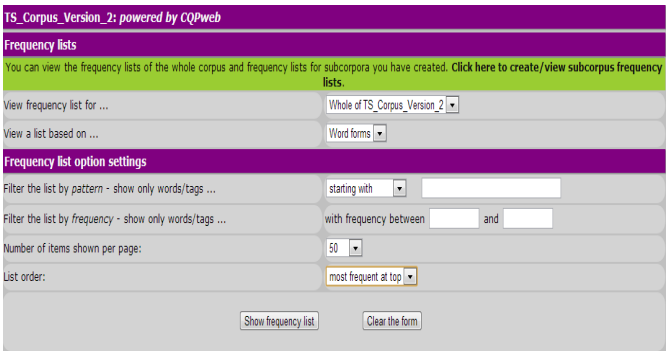
Tablo-5: Kelime Arama (Word look up)



Bu arayüzde sorgulamak istediğimiz "sözcük/biçimbirim/ses ile başlayan (starting with), biten (ending with), devam eden ya da tam olarak o yapıyı karşılayan" sonuçları sözcük türü etiketi ile birlikte ya da sözcük türü etiketi olmadan elde edebilmekteyiz.

"Sıklık Listeleri (frequency lists) menüsünden (Tablo-6), Tablo-5'te gördüğümüz arama seçeneklerine bağlı olarak sözcüğün kök şeklinde kullanım sıklığı da dâhil olmak üzere kök ve biçimbirim bazında kullanım sıklığını elde edebilmekteyiz. Ancak herhangi bir biçimbirimin sıklığını denetlemek için "sözcük biçimleri (word forms)" seçeneğine bağlı olarak "ile bitenler (ending with)" ayarı ile sorgulatmamız gerekmektedir. Bu şekilde biçimbirim sorgulandığında sonuçlar listesinde sorgulanan biçimbirimle sesteş olan yapıları da veriyor olması derlemin bir kusurudur.

Tablo-6: Sıklık Listeleri (Frequency List)



"Anahtar sözcükler (keywords)" menüsü kullanıcının oluşturduğu alt derlemlerdeki sıklık listelerini karşılaştırma imkanı veren bir uygulamadır.

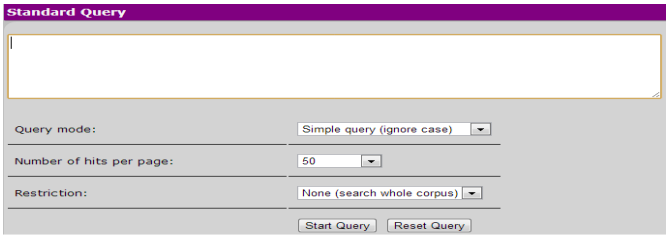
"Kullanıcı kontrol (user controls)" menüsünde ise "kullanıcı ayarları (user setting), sorgu geçmişi (query history), kaydedilmiş sorgular (saved queries), kategorileştirilmiş sorgular (Categorised queries), bilgisayara kaydedilmiş bir sorgu sonuçlarını kaydetme (upload a query) ve alt derlem oluşturma (Create/edit subcorpora)" seçenekleri (bkz. Tablo-4) bulunmaktadır.

Kullanıcı kontrol menüsünün altında Ts Corpus ve CQPweb hakkında bilgi veren menü (bkz. Tablo-4) bulunmaktadır.

"Ts Corpus"un "menü" kısmı hakkında bilgi verdikten sonra şimdi derlemde yapabileceğimiz sorgulama biçimlerine değinmek faydalı olacaktır.

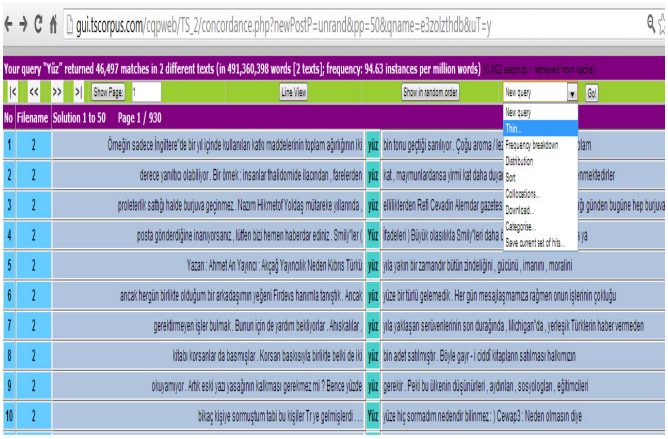
Sınırlandırılmış arama (restricted query), sözcük bakma (word lookup) ve sıklık listeleri hakkında yukarıda bahsetmiştik. Sınırlandırılmış aramaya da uygulayabileceğimiz standart sorgu arayüzündeki (bkz. Tablo-7) sorgu şekilleri hakkında şunları söyleyebiliriz.

Tablo-7: Standart Sorgu



Sorgu arayüzüne "yüz" sözcüğü girildiğinde büyük/küçük harfe duyarlı olamayan basit arama [Simple query (ignore case)] ile tam olarak "yüz" sözcüğüyle örtüşen 46.497 sonuç (bkz. Tablo-9) elde edilecektir. Bu aramayı büyük/küçük harfe duyarlı (simple query case-sensitive) şeklinde yaptığımızda ise 41.656 küçük, 4.413 büyük harfle başlayan sonuç elde edilecektir. Büyük/küçük harfe duyarlılık sorgu arayüzüne sözcük nasıl girildiyse o sözcük baz alınarak yapılacaktır.

Tablo-9: Sonuç Arayüzü



Tablo-9'da gördüğümüz "yeni sorgu (new query)" menüsünden sonuçlarla ilgili şu eylemleri yapabiliriz:

-Sonuçları daraltabiliriz (thin)

-Sonuçların frekansını alabiliriz (frequency breakdown)

-Genel veri ve gazete verileri olmak üzere iki dosyadan oluşan derlemde sözcüğün dağılımını (distribution) görebiliriz

-Sözcüğün sağında ve solunda bulunan sözcüklere göre kısaltma (sort) yapabiliriz

-Sorguladığımız sözcüğün eşdizimliliğini (collacations) görebiliriz.

-Sonuçları bilgisayarımıza indirebiliriz

-Sonuçları sınıflayabiliriz.

-Mevcut kümeyi ya da seçilmiş bir kümeyi kayıt edebiliriz (save current set or hits…)

Sözcük türü yönünden etiketlenmiş olan derlemde sorgu arayüzüne "etiket" kodlarını girerek sözcük türü şeklinde sorgu yapmak mümkündür. Sorgu arayüzüne _Verb yazdığımızda derlemde fiil olarak işaretlenen tüm sonuçlar elde edilecektir[10]. Bu şekilde yapılan aramalarda fiil kökünden türeyen isimlerin de fiil olarak verilmesi doğru değildir.

Sorgu arayüzünde sözcüğün kökü (lemma), {KÖK} şeklinde girildiğinde sözcüğün biçimbirim almış/almamış kullanımları elde edilmektedir. Bu tür sorgu yapmanın iki faydası vardır. Bu faydalardan biri {gönül} sözcüğünü sorguladığımızda bu sözcüğe –Im, -In vb. biçimbirimlerinden

[10] TS Corpus sözcük türü etiketleri http://tscorpus.com/tr/faq/detay/sozcuk-turu-etiketi adresinde verilmiştir.

birinin eklenmesi sonucu orta hecesi düşmüş olan "gönlüm, gönlün" gibi sözcükbirimleri de getiriyor olmasıdır. Bir diğer faydası ise "p,ç,t,k" sesleriyle biten sözcüklere ünlü ile başlayan bir biçimbirim geldiğinde "b,c,d,ğ" seslerine dönüşen yapıları da vermesidir.

Derlemdeki verilerin biçimbirim olarak işaretlenmiş olması dolayısıyla TS Corpus biçimbirim olarak sorgu yapma imkânı tanımaktadır. Sorgu arayüzüne [Moph=biçimbirim etiketi] girilerek işaretlenmiş olan bir biçimbirimin kullanım sonucu elde edilebilmektedir. Örneğin sorgu arayüzüne –mAK isim-fiil biçimbiriminin etiketi olan [Morph=".*\Inf1\+.*"] girildiğinde –mAK isim-fiil biçimbiriminin bulunduğu 3.235.358 sonuç elde edilecektir.

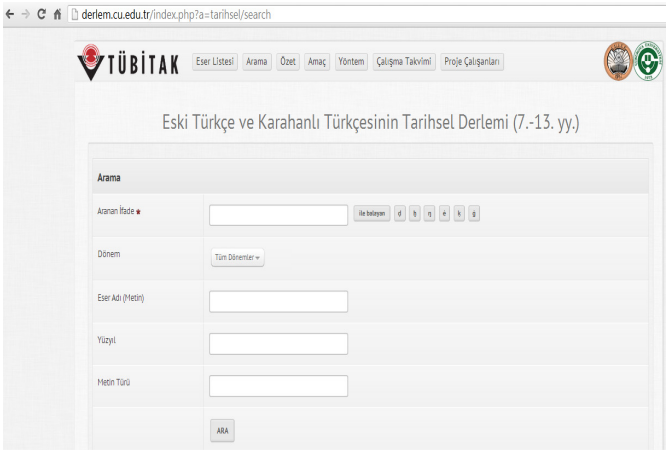
Ts Corpus kullanıcılara *, ?, +, @, /, ( ), [ ], -, _, < > gibi joker işaretleri ile arama imkanı sunmaktadır.

Çalışmamızda üçüncü çevirimiçi derlem olarak değineceğimiz çalışma Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D)'dir. Türkçenin artsüremli/tarihsel derlemi olan Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D) Orhon Türkçesi, Uygur Türkçesi ve Karahanlı Türkçesi'ne ait yazılı metinlerin elektronik ortama aktarılarak sözcükbirim ve sözdizim bazında işaretlenmesiyle oluşturulmuş 600 yıllık bir dönemi (7-13. yy) kapsayan 400-450 bin sözcük içeren çevirimiçi derlemdir[11].

Derlemin sorgu sayfasına http://derlem.cu.edu.tr/index.php?a=tarihsel/search adresinden erişilebilmektedir. Derleme giriş yapmak için herhangi bir kullanıcı adı ya da şifreye ihtiyaç yoktur.

Derlemin sorgu sayfası Tablo-10'da gösterildiği gibidir.

Tablo-10



Karşımıza gelen ekranda sorgu arayüzünün sağ tarafında (*) "ile başlayan" joker ve "ḍ, ḫ, ė, Ŋ, ġ, ķ" karakterleri bulunmaktadır. Sözcük sorgularımızı "Karahanlı Türkçesi, Uygur Türkçesi ve Orhon Türkçesi" dönemlerinin tümünü ya da bu dönemlerden birini kapsayacak şekilde ve "eser adı (metin), yüzyıl ve metin türü" bağlamında sınırlandırarak yapabilmekteyiz.

[11] http://derlem.cu.edu.tr/index.php?a=tarihsel/icerik_amac, (çevirimiçi) 02.04.2013

Standart şekilde sorgu arayüzüne yazacağımız herhangi bir sözcükle (ıgaç) arama yaptığımızda Tablo-11'de gösterilen sonuç menüsü ekrana gelmektedir.

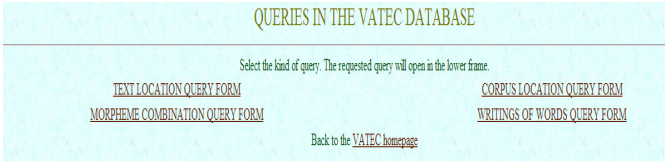
Tablo-11

Sonuç sayfasında sorgulattığımız sözcük, sözdizim bağlamında metin türü, yüz yılı, yer aldığı eserin adı ve dönemi bilgilerinin de verildiği bir yapıda sunulmaktadır. Sözcüğün geçtiği cümlenin ait olduğu eserdeki beyit/satır numarası cümlenin sol tarafında verilmektedir.

Joker karakteri ile herhangi bir sözcük ve bu sözcüğün bçimbirimli kullanımlarına ulaşabilmekteyiz. Örneğin sorgu arayüzüne "ıgaç*" şeklinde sözcüğü girdiğimizde Tablo-11'deki sonuçlara ek olarak "ıgaçka" sözcüğü de elde edilecektir.

Son olarak vereceğimiz çevirimiçi derlem "Vorislamische Alttürkische Texte: Elektronisches Corpus 'VATEC' (İslamiyet Öncesi Türkçe Metinleri Elektronik Derlem)"i 1999-2003 yılları arasında Deutsche Forschungsgemeinschaft tarafından desteklenen ve Prof. Dr. Marcel Erdal başkanlığında gerçekleştirilen bir projedir[12]. Derlem web sayfası Almanca olarak sunulmuştur. Derleme http://vatec2.fkidg1.uni-frankfurt.de/ adresinden ulaşılmaktadır. "VATEC veritabanı için çevirimiçi arama motorları" bağlantısından derlem sorgu arayüzleri seçeneklerine (bkz. Tablo-12) bağlanılmaktadır[13].

Tablo-12



"Metin yeri sorgu şekli (Text location query form)" sözcükleri tüm kategorilerde veritabanında arama imkânı vererek dil, metin ve kategori yönünden sorguyu kısıtlama olanağı vermektedir.

VATEC veritabanında Eski Türkçe sözcükleri arama olanağı sunan "Derlem içi arama şekli (Corpus Location query form)" tam olarak aranan sözcüğü sorgulatma yanında "*" joker karakter sayesinde aranan sözcük ile başlayan diğer

[12] http://vatec2.fkidg1.uni-frankfurt.de/ (çevirimiçi) 10.04.2013

[13] http://vatec2.fkidg1.uni-frankfurt.de/vatecasp/query.htm Bu sayfadaki sorgu seçenekleri ile ilgili açıklamalardan (explanations) faydalanılmıştır.

sözcükleri bulma imkânı da sunmaktadır. Sonuç sayfasından da sözcüğün geçtiği bağlama ulaşılmaktadır.

Sözcükbirim kombinasyonu arama şekli (Morpheme combination query form) menüsünde, sözcükbirim/biçimbirim yönünden etiketlenen Eski Türkçe metinlerde sözcükbirim ve/veya bçimbirim sorgulanabilmektedir.

İslamiyet öncesi Göktürk (Runik), Uygur, Mani, Tibet, Çin, Süryani ve Brahmi alfabeleriyle yazılan Eski Türkçe sözcüklerin yazı sistemlerindeki dağılımına "Sözcüklerin yazı sistemleri sorgu şekli (Writtings of words qoery form) menüsünden ulaşılmaktadır.

Bu sorgu seçeneklerinden "derlem içi arama şekli" sorgu arayüzüne "bašlayu" (başlayarak) sözcüğünü girerek standart bir arama yaptığımızda Tablo-13'teki gibi sonuçlar elde edilecektir[14].

Tablo-13: Vatec Derlem İçi Arama Şekli Sonuç Sayfası

These are the results for your query:

TRANSCRIPTION	PRECISE TRANSCRIPTION	TRANSLITERATION	WRITING SYSTEM
LANGUAGE	NAME OF TEXT (click on name leads to context)	SECTION OF TEXT	LINE OF TEXT (REFERENCE)
	MORPHEME (if analyzed)	GLOSS	PART OF SPEECH / GRAMMATICAL FUNCTION
bašlayu	bašlayu	bšlyw	Manichaean
Old Turkic	Manichäisch-Türkische Texte	MT0941-0948	MT942(M657:r2)
	bašla	beginnen	vt
	-yU	-GERA	-gerund

Sonuç sayfasında sözcüğün geçtiği metnin ismini (name of text) tıkladığımızda sözcüğün kullanıldığı bağlama erişilmektedir[15].

Tablo-14

Referenz:

MT942(M657:r2)

Transliteration	swyrwgwm	.		ym	bšlyw	
genaue Transkription	sürügüm	..	&	ymä	bašlayu	
Transkription	sürügüm	..		ymä	bašlayu	
Morphem	sürüg	-(X)m	..	ymä	bašla	-yU
Glossierung	Herde	-POSS1	..	nun	beginnen	-GERA
Wortart/Funktion	n	-possessor	punct	conj	vt	-gerund

Übersetzung: Herde. Nun beginnend

Kommentar: Man. Interpunktion: ein scharzer Punkt in einem roten Kreis.

4. Sonuç

Türkçe çevirimiçi derlemler üzerine yaptığımız çalışma neticesinde yazılı metinlerin temel alındığı dört derlem tespit edilmiştir.

Bu derlemlerden Türkçe Ulusal Derlem %95'i yazılı metinlerden oluşan dengeli-genel derlemdir. Derlem sözcükbirim olarak etiketlenmiştir. Joker karakterler sayesinde kullanım kolaylığı sağlanmıştır. Yayın yılı, medya, yazar cinsiyeti gibi çeşitli sınırlamalar ile arama yapılabilmektedir.

Standart ya da sınırlandırılmış sorgu neticesinde derlem içerisinde sözcüğün kaç defa kullanıldığı ve frekansı verilerek, kullanıldığı bağlama ulaşılmaktadır.

TUD'un sonuç sayfasındaki "dağılım, listele ve eşdizimlilik" seçenekleri vasıtasıyla sözcüğün "yayın yılı, medya, metin örneklemi, alan, türev metin biçimi, yazarın cinsiyeti, yazar/yazarların türü, okuyucu kitlesi ve tür" yönünden dağılımını; "listele" seçeneği ile anahtar sözcüğün sağ ve sol tarafında bulunan sözcüklerin –istenildiğinde alfabetik sıraya göre- birer sütun oluşturularak dizilimini ve "eşdizimliği" ile anahtar sözcüğün sağ ve/veya sol tarafında bulunan sözcüklerle kullanım sıklığını görebilmekteyiz.

Biçimbirim yönünden etiketleme yapılmadığı için (*) karakteri kullanılarak yapılan sorguda, istenilen biçimbirimin dışında bu biçimbirimle sesteş olan sözcükbirimler de sonuç listesinde sıralanmaktadır.

Ts Corpus sanal ortamdaki gazete, form, sohbet gibi yazışmaların sözcükbirim, kök ve biçimbirim olarak etiketlenmesi ile hazırlanmış genel-dengesiz bir derlemdir. Derlem, belli kurallar çerçevesinde oluşturulan yazılı metinlerde az karşılaşılacak daha çok konuşma diline ait sözcükbirim ve biçimbirimlerin bulunması yönüyle önemlidir.

Standart ve sınırlandırılmış arama yapılabilinen Ts Corpus'ta joker karakterleriyle de arama yapılmaktadır. Aranan sözcüğün eşdizimliliği, derlem içindeki dağılımı, sonuçların daraltılması ve kayıt edilebilmesi ve sözcüğün kullanıldığı bağlama erişilebilmesi derlemin belli başlı özellikleridir.

Orhon, Uygur ve Karahanlı Türkçesi eserlerinin önemli bir kısmını bir araya getiren ETKT-D'i sözdizim ve sözcükbirim bağlamında etiketlenmiştir. '*' karakteri ile arama yapılabilmektedir. Sonuç sayfasında aranan sözcüğün dönem, yüzyıl, metin türü ve hangi eserde geçtiği bilgilerine ek olarak sözcüğün eserdeki beyit/satır numarası da verilmektedir.

Uygur Türkçesi metinlerinin önemli bir kısmının sözcükbirim, kök ve biçimbirim şeklinde etiketlendiği VATEC, aranan öğenin geçtiği metni, türünü, anlamını ve işlevini vermektedir.

Kaynakça

Aksan, Yeşim ve Mustafa Aksan. Building a national corpus of Turkish: Design and implementation. Working Papers in corpus-based linguistics and language education No:3, 299-310 Tokyo: Tokyo University of Foreign Studies, 2009.

Burkhanov, Igor. Lexicography A Dictionary of Basic Terminology, Rzeszów: Wyższa Szkoła Pedagogiczna, 1998.

Eski Türkçe ve Karahanlı Türkçesinin Tarihsel Derlemi (ETKT-D),

[14] http://vatec2.fkidg1.uni-frankfurt.de/vatecasp/query.htm (çevirimiçi) 10.04.2013

[15] http://vatec2.fkidg1.uni-frankfurt.de/vatecasp/Manich%C3%A4isch-T%C3%BCrkische_Texte.htm#208628 (çevirimiçi) 10.04.2013

http://derlem.cu.edu.tr/index.php?a=tarihsel/search 30.03.2013.

Ercilasun, A. Bican, Başlangıçtan Yirminci Yüzyıla Türk Dili Tarihi, Akçağ Yay., Ankara, 2012.

Halliday, M. A. K. vd., Lexicology and Corpus Linguistics (Open Linguistics), London: Contiuum, 2004.

Kennedy, Graeme, An Introduction to Corpus Linguistics, Longman, 1998.

McEnery, T., A., Wilson, Corpus Linguistics An Introduction, Second Edition, Edinburgh: Edinburgh University Press, 2001.

McEnery, T., A., Hardie, Corpus Linguistics: Metod, Theory and Practice, Cambridge: Cambridge University Press, 2012.

Say, Bilge, vd. "Development of a corpus and a treebank for present-day written Turkish." Proceedings of the eleventh international conference of Turkish linguistics. 2002.

Sözlü Türkçe Derlemi (STD), http://stc.org.tr/hakkinda/ 30.03.2013.

TS Corpus, http://www.tscorpus.com/tr, 30.03.2013.

Vorislamische Alttürkische Texte: Elektronisches Corpus 'VATEC', http://vatec2.fkidg1.uni-frankfurt.de/ 30.03.2013.

Türkçe Ulusal Derlem (TUD) http://www.tnc.org.tr/ 30.03.2013