Derlem dilbilim yöntemlerinin etkin olarak araştırmalarda kullanımı: Uygulamalar

Şükriye RUHİ, Mustafa AKSAN, Yeşim AKSAN[*]

Özet

Türkçe yazılı ve sözlü derlem tasarımlarının son on yılda yoğunlaşmış ve kullanıma açılmış olması, Türkçe derlem dilbilimsel araştırma yöntemlerinin tartışılması ve derlemlerin etkin kullanımlarının yaygınlaşmasını gerekli kılmaktadır. Bu yuvarlak masada sunulan bildiriler ve tartışmalarda, güvenilir ve tekrarlanabilir nicel ve nitel araştırma yöntemleri, derlem oluşturma ve kullanmanın etik kuralları, Türkçe derlem araştırmalarını bekleyen yeni çalışma ve eylem planları ele alınmıştır.

Anahtar sözcükler: derlem dilbilimi araştırma yöntemleri, istatistiksel çözümlemeler, nitel araştırmalar, Sözlü Türkçe Derlemi (STD), Türkçe Ulusal Derlemi (TUD)

1. Giriş[1]

Son on yılda Türkçe yazılı ve sözlü derlem tasarımları yoğunlaşmış ve kullanıma açılmıştır (bkz. Aksan, Aksan, Koltuksuz ve diğerleri 2012; Ruhi, Eröz-Tuğa, Hatipoğlu, Işık-Güler, Acar, Eryılmaz ve diğerleri 2010; Say, Zeyrek, Oflazer ve Özge, 2002). Derlemlerin araştırmalarda kullanımı bazı başka alanlarla yöntemsel benzerlik gösterse de kendine özgü yöntemleri olduğu bir gerçektir (bkz. Baroni ve Evert 2009; Oakes 1998). Öte yandan sadece kendi içindeki yöntem ve araç geliştirme gereksinimleri nedeniyle değil (bkz. Leech, 2011) derlem dilbiliminin başka alanlarla kesişmesi nedeniyle de derlem dilbiliminde yöntem geliştirme günümüzde çok önemli bir tartışma alanı oluşturmaktadır (bkz. örn. Gries, 2013). Bu bakımlardan hem Türkçenin derlem araştırmalarının sayıca az olması hem de alanın dilbilim topluluğu içinde oldukça yeni olması varsayımlarından yol çıkarak, derlem dilbilimi araştırma yöntemlerinin tartışılması gerektiği düşünülmüştür. Bu amaçla Sözlü Türkçe Derlemi (STD; http://stc.org.tr) ile Türkçe Ulusal Derlemi'nden (TUD; http://www.tnc.org.tr) örneklerle bir yuvarlak masa toplantısı 27. Ulusal Dilbilim Kurultayı'nda düzenlenmiştir. Yuvarlak masada sunulan bildirilerin arka planını sunmak amacıyla bu yazıda derlem tanımından başlayarak derlemlerin araştırmalara genel etkisi ve bazı genel geçer uygulama kurallarından kısaca söz edilecektir. Yazının son bölümünde Türkçe derlem dilbiliminin gelişmesi amacıyla yürütülen bazı çalışmalar ve araştırma eylem planları konu edilecektir. Yuvarlak masa bildirilerinin derlemlerin daha etkin bir biçimde kullanılması konusuna katkıda bulunacağını umuyoruz.[2]

2. Derlem ve derlem araştırmaları

2.1 Derlem nedir?

Derlem araştırma yöntemlerinin çerçevelerini kullanılan verilerin özellikleri bakımından çizmek amacıyla önce derlem tanımı üzerinde duracağız. Özellikle genel amaçlı kullanım için oluşturulması derlemlerden (İng. general/reference corpus) söz edersek, derlem terimi başka dil kaynaklarının bazı niteliklerini paylaşmakla birlikte olmazsa olmaz başka nitelikler taşımaktadır. Söz konusu niteliklerin ne olduğundan söz edelim. Belli bir dilin veya dil değişkesini temsil edebilme amacıyla, belli bir zaman aralığında yazılı ve/veya sözlü dil kullanım metinlerini/konuşmalarını, coğrafi bölge, yazar/konuşan özellikleri (cinsiyet, yaş, eğitim vb.) ve iletişim ortamlarının alan ve türlerine (İng. domain, genre) göre dengeli ve katmanlı örnekleme yoluyla derleyip, ayrıntılı veribilgisi (İng. metadata) ve temel dilbilimsel çözümlemelerle birlikte elektronik ortamlarda sunan kaynaklara derlem denir (Biber, Conrad ve Reppen, 1998; Gries, 2009; McEnery ve Hardie, 2012). Bu bakımdan derlemler zaman aralığı seçmeleri nedeniyle arşiv gibidirler, ama arşivler söz konusu denge ve katmanlı örnekleme yoluyla oluşturulmazlar ve doğal iletişim metinleri içermeyebilirler. Sadece belli konu ve sözcük örneklerinden oluşmadıkları için derlemler indeks veya sözcük listeleri değildir. Metin ve konuşmalar veribilgileriyle birlikte dilbilimsel çözümleme içerdiği için derlem ham bir veritabanı değildir. İnternet tarayıcıları aracıyla WWW'den

[*]Prof.Dr. sukruh@metu.edu.tr; sukriyeruhi@gmail.com; Prof. Dr., Mersin Üniversitesi, mustaksan@gmail.com, Prof.Dr. Mersin Üniversitesi, yesim.aksan@gmail.com

[1] Bu yazı, düzenlenen yuvarlak masanın giriş ve sonuç tartışmalarının kısaltılmış halidir.

[2] Yuvarlak masa tartışmaları yöntem konusuna odaklanmakla birlikte, derlem dilbiliminin sadece dil inceleme yöntemlerinden ibaret olmadığını, verilere dayalı, tümevarımlarla "dil incelemelerine yeni bir felsefi yaklaşım" (Tognini-Bonelli, 2001: 1) getirdiği ve dil hakkındaki düşünceleri gözden geçirmeyi gerekli kıldığını (Sinclair, 1994) belirtmek isteriz. Derlem dilbilimini ağırlıklı olarak bir inceleme yöntemi gören çalışmalar için bkz. McEnery, Xiao ve Tono (2006).

edinilen metin örneklemleri her ne kadar yeni iletişim ortamları, yeni sözcükler vb. dil kullanımları hakkında bilgi sağlasalar da aynı nedenle derlem değildir.[3] Son olarak, derlem arayüz araçları derlem değildir.

2.2 Derlemlerin dilbilimsel araştırmalara etkisi

Yukarıdaki tanım üzerinden gidersek, derlem tasarımından derlem incelemesine kadar derlem araştırmaları kuramsal çalışmalardır. Sadece derlem temsiliyeti (İng. representativeness) kavramını alırsak, herhangi bir derlemin temsiliyeti farklı ölçütlerle geliştirilmiş olabilir. Örneğin, metin seçimi hem yazılı hem de sözlü derlemlerde katmanların üretici – dinleyici/okur – metin ilişkisine göre mi yapılacağı ya da dil kullanım özelliklerinin mümkün olan en geniş biçimde temsil edilmesine göre mi yapılacağı (Leech, 2007) derlemlerin dilsel özelliklerini önemli ölçüde etkilemektedir. Bu bakımdan farklı ölçütlere göre tasarlanmış derlemler farklı nicel ve nitel sonuçlar içerecektir; dolayısıyla da dilin veya değişkenin farklı betimlenmesine yol açacaktır. Örneğin, Çek Ulusal Derlemi sözlü bileşeninde göre coğrafi bölgeye katmanlı bir tasarım kullanmaktadır; ancak konuşanların cinsiyet, yaş ve eğitim temsiliyetleri nüfustaki oranlara göre değil eşit temsiliyete göre tasarlanmıştır (Válková, Waclawičová ve Křen, 2012, s. 3347). Dolayısıyla bu derlem üzerinden yapılacak dil kullanımı-konuşan özellikleri üzerine yapılacak araştırmalarda genelleme söz konusu olduğunda dilsel öğelerin kullanım sıklıklarından söz ederken nüfusa göre bir betimlemeden söz edilemez. Bu demektir ki derlemlerin özellikleri hangi soruların sorulabileceğini ve araştırmanın nasıl tasarlanacağını yakından etkiler (bkz. Biber ve Jones, 2009).

Temsiliyet kavramı ile örneklendirdiğimiz konu, hem derlem tasarımında hem de derlem araştırmalarında veritabanı olarak kullanılan derlemin nitelikleri hakkındaki belgelerde şeffaflığı zorunlu kılar. Bu konunun önemini Sinclair (2005, s. 98) şu şekilde açıklamaktadır:

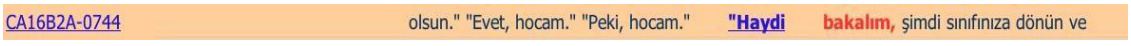
> Derlem kuranlar, kendi ortamlarında kurabilecekleri en iyi derlemi kurarlar. Burada en doğru tutum derlemin içeriği konusunda ayrıntılı ve dürüst olmalarıdır. Derlem kuranın derlemini nasıl betimlediğine bakarak, derlemi kullanan araştırmacılar vardıkları sonuçların ne kadar güvenilir olacağını ölçebilirler, aynı derlemi gelecekte kullanacak olanlar da, kendi amaçları açısından ne ölçüde güvenilir olduğunu değerlendirebilirler.

Şeffaflık aşağıda açıklanacağı üzere derlem araştırmalarında başka alanları da ilgilendirmektedir.

2.3 Derlem araştırmalarında şeffaflık ve bazı temel etik uygulamalar

Tüm bilimsel araştırmalarda olduğu gibi, derlem araştırmalarının da etik kuralları gözeten uygulamaları bulunmaktadır. Ancak elektronik derlemlerin oluşturduğu veriler belli kullanım kuralları çerçevesinde geniş araştırma topluluklarınca paylaşıldıkları için kaynakları oluşturan kaynaklara girdi sağlayan kişilerin fikri mülkiyet hakları ve verilerin gizliliğinin korunması konuları başlı başına bir araştırma alanı oluşturmaktadır (bkz. McEnery ve Hardie, 2012; Rehm ve Witt, 2007). Derlem oluşturma ve derlem araştırması yapmak birçok etik kuralın uygulanmasını gerektirmektedir.

Yukarıda şeffaflıktan söz etmiştik. Derlem araştırmalarının diğer araştırma türlerine göre bir önemli üstünlüğü aynı çalışmanın başkaları tarafından tekrarlanabilir olmasında ve böylelikle çeşitli çalışmalarla varılan sonuçların sınanabilmesinden kaynaklanmaktadır. Bunu sağlamak amacıyla derlem kullanıldığında alıntılarda veya bağımlı dizin örneklerinde dosya künyesi belirtmek zorunludur (bkz. Resim 1).



Resim 1. TUD-Tanıtım Sürümü bağımlı dizinde dosya künyesi

Aynı şeffaflık kullanılan derlemi url adresi ve derlem kurucuların belirlediği yayınları kaynak göstermekle de sağlanır. Öte yandan yapılan araştırmada kullanılan açık kaynak yazılımlar kaynakçada gösterilmeli ve bunlara yapılan değişiklikler yazılımın kullanım ilkelerine uymalıdır (McEnery ve Hardie, 2012, ss.60-69).

Yukarıda anılan kurallar kadar hassas olan bir konu ise metin alıntılarını ilgilendirmektedir. Derlemlerden yapılacak alıntıların uzunluğu, içerdikleri bilgi vb. konularda derlem kurucularının kullanım sözleşmelerinde belirttikleri kurallara uymak derlem yayıncılığının sağlıklı sürdürülebilmesi için zorunlu olduğu gibi, derleme veri sağlayan kişilerin haklarını da korur. Örneğin, sözlü derlemlerde konuşanların kimliğini ortaya çıkarabilecek bilgilerin yazılı veya sözlü olarak yayınlanmasına kısıtlamalar getirildiği gibi, konuşanları incitebilecek içerikte

[3] Karşıt bir görüş için bkz. Kilgarriff, A. ve Grefenstette (2003) ve Leech (2007).

yayın yapılması etik değildir (örn. STD için bkz. http://stc.org.tr/wp/wp-content/uploads/2012/11/kullanici_anlasmasi_tanitim_surumu2.pdf).[4]

Bu yazı çerçevesinde elbette derlem araştırmalarının ancak bazı özellikleri üzerinde durulabildi. Ruhi (bu kitap) ve Aksan ve Demirhan'da (bu kitap) derlem dilbilimi araştırmalarında nitel ve nicel yöntemlerin çeşitli yönleri üzerinde durulmaktadır.

3. Türkçe derlem dilbilimi araştırmalarında çalışmalar ve gereksinimler

6-7 Ekim 2011 tarihleri arasında düzenlenen "Ulusal Konuşma ve Dil Teknolojiler Platformu Kuruluşu: Türkçe'de Mevcut Durum" çalıştayında (Doğan, 2011) derlem dilbilimini yakından ilgilendiren şu görüşlere yer vermiştir:

- Türkçe veri ve kaynakların yetersizliği
- Veri setlerinin yetersizliği
- Türkçe dil derlemlerinin yetersizliği
- Son kullanıcı için ürünlerin yetersizliği

Bu eksikliklerin yanı sıra Türkçe derlem dilbilimi için ayrıca önemli gördüğümüz ve üzerinde çalıştığımız konular şunlardır:

- Konuşma Türkçesinin (ağız) derlemlerinin geliştirilmesi
- Sözlü ve yazılı Türkçe derlemlerde standart belirlenmesi ve bunların TEI gibi uluslararası standartlara uyumu
- Derlem oluşturma, işleme, yayınlama ve kullanım ilkelerinin belirlenmesi
- Veribilgisi ve açımlama için ontolojilerin geliştirilmesi
- Türkçe derlemlerde işaretleme ve açımlama çalışmalarının geliştirilmesi
- Türkçe çözümlemeler için yazılımların değerlendirilmesi ve uyarlama çalışmalarının yapılması
- Türkçe derlemler için portal geliştirilmesi

Aşağıdaki bölümlerde bu çalışmaların bazı yönleri üzerinde durulmuştur.

Derlem kaynakları geliştirme ve ilintili çalışmalar

Hali hazırda doğal ortamlarda konuşulan Türkiye Türkçesi bildiğimiz kadarıyla STD ve TUD'da temsil edilmektedir. Her iki derlemde Türkçenin ağızları temsil edilse de söz konusu kaynaklar ağız derlemleri değildir. Bu bakımlardan konuşma Türkçesinin hem ölçünlü hem de ağız özelliklerinin karşılaştırmalı olarak incelenebilmesi için başka derlemlerin tasarlanarak yayınlanması gerekmektedir. Türkiye'de Türkoloji geleneğinde ağız araştırmaları için veri toplanmaktadır; ancak söz konusu veriler dil araştırmaları topluluklarına sunulmamaktadır. Keza, çeşitli söylem çözümlemesi ve edimblim araştırmaları için konuşma verilerin ilk veri toplama koşullarının el verdiği ölçüde derlem formatlarında yayınlanabilmesi Türkçe araştırmalarının veri setlerini zenginleştirecektir.[5] Çok emek yoğun olan ve parasal kaynak gerektiren bu çalışmaların ürünlerinin derlem olarak Türkçe üzerine araştırma yapan bilim insanlarının kullanımına sunulmaması ve araştırma desteği sağlayan kurumların derlem veya veri setleri için yayınlanma ölçütü kullanmaması (ya da en azından teşvik etmemesi) zaten zayıf olan disiplinlerarası araştırmaların önünde bir engeldir. Bu bağlamda Türkiye Türkçesi derlemi olmamakla birlikte, ODTÜ-KKK Kıbrıs Türkçesi Derlemi projesi çerçevesinde geliştirilmekte olan Kıbrıs Türkçesi derlemini (STCDC) anmak yerinde olur (bkz. Ruhi ve Işık-Taş, basımda). STCDC, STD ile aynı derlem tasarımı ilkeleri ve derlem oluşturma yazılımlarını kullanmakta ve birbirine çok yakın çeviriyazı ilkeleri gözetmektedir. Bu bakımdan yayınlandığında karşılaştırmalı ağız araştırmalarına yeni bir bakış açısı ve yöntem sunabilecektir. Bu konu bizi yukarıda anılan veribilgisi çalışmalarına getirmektedir.

Derlemlerde ortak veribilgisi sistemlerinin geliştirilmesi ve kullanılması derlemler araştırmalarının karşılaştırmalı olarak yürütülmesini sağlamaktadır. Her ne kadar derlemler benzerlik gösteren veribilgisi terimleri kullansalar da, veribilgilerinin bağlantılı olduğu nesneler arasındaki ilişkiler farklı tasarlanmış olabilir (Hedeland ve Wörner, 2012). Bu bakımdan hem terimler arası ortaklık hem de veribilgisi yapıları arasında ortak paydaların sağlanmasını kolaylaştıracak ontolojilerin geliştirilmesi çok önemlidir. Aynı şey sözlü derlemlerde

[4] Derlem araştırmalarındaki başka etik uygulamalar için bkz. Adolphs ve Knight (2010). Derlemler araştırmalarındaki yasal konular için bkz. Rehm ve Witt (2007).

[5] Bu tür büyük oylumlu çalışmalar için bkz. Australian National Corpus (Avustralya Ulusal Derlem, Cassidy, Haugh, Peters ve Fallu, 2012) ve Datenbank für Gesprochenes Deutsch, DGD2 (Sözlü Almanca Veri Bankası, http://dgd.ids-mannheim.de). Kanımızca bu tür çalışmalar önümüzdeki yıllarda dil kaynaklarını geliştirmede daha da önem kazanacaktır.

çeviriyazı ölçünleştirme ve ontolojileri çalışmaları için de geçerlidir (Cassidy, 2013; Schmidt, basımda). STD ve yukarıda sözü edilen Kıbrıs Türkçesi Derlemi araştırmaları kapsamında söz konusu ortaklıklar sağlanmıştır ve derlem veri formatları büyük ölçüde TEI ile uyumludur (bkz. Schmidt, 2011). Ancak konuşma derlemlerinin sayısının artmasıyla veribilgisi ve çeviriyazı standartlarının sınanması gerekmektedir. Söz konusu gelişmeler derlemlerin ve veri setlerinin sürdürülebilirliğini güçlendirecektir.

Ölçünleştirme çalışmaları sadece veri formatları, veribilgileri ve çeviriyazıyı ilgilendirmemektedir. Yazılı metinler bağlamında derlem tasarımı oldukça yerleşik bir uygulamaya varmış görünse de, günümüzde elektronik iletişimin türlerinin artmasıyla birlikte derlemlerin alan ve tür bakımlarından gözden geçirilmesi gerekmektedir. Sözlü derlemlerdeki durum daha da zorlayıcıdır; çünkü bugüne kadar dengelilik ve temsiliyet ölçütleri üzerinde tartışmalar hala sürdüğü gibi neredeyse tüm diller için belirsiz olan bir durum söz konusudur; şöyle ki, konuşanların ne tür sözlü dil kullanımlarına konuşan ve/veya dinleyici olarak katıldıklarının tespiti bir araştırma sorunudur. STD, Çek Ulusal Derlemi'nin yazılı bileşeni için izlediği yöntemi konuşma dili için kullanmıştır (bkz. Leech, 2007, s. 147). Buna göre bir pilot çalışmayla küçük bir anadili konuşanı topluluğuna çeşitli konuşma ortamlarına katılma sürelerini günlüklerle tespit etmeleri istenmiştir. Bu tür yöntemlere rağmen ortam türleri veri toplama aşamalarında geliştirilmek durumundur kalınmıştır. Örneğin, aile içi ve işyeri iletişim ortamlarına görüntülü konuşma bir tür olarak eklenmiştir.

Derlem oluşturma araştırmalarında işaretleme ve açımlamaların geliştirilmesi önemli bir çalışma alanıdır. Bu bölümde sadece konuşma dilinin işaretleme ve açımlamaları üzerinde duracağız. Bildiğimiz kadarıyla sözlü Türkçe için biçimbirim işaretlemesi günümüze kadar yapılmamıştır. STD'deki çalışmalardan ve buradan edindiğimiz deneyimden söz edersek, konuşma Türkçesine özgü sözlüksel olmayan birimler (örn. *hı-hı* ve *hmm* gibi geribildirimler) biçimbirim işaretleme araçlarının sözlüklerine eklenebilir. Ancak çok sözcüklü birim oluşturan yansımalı birimlerin (örn. tiyatro miyatro) ve kısaltmaların tümünü önceden kestirmek güç olmaktadır. STD'nin ikinci sürümünde için yapılan çalışmalarda bu gibi nedenlerden ötürü biçimbirim çözümlemeleri hizmet alımı ortamlarına kısıtlanarak sürdürülmektedir.

Açımlama konusuna sözce birimlerinin tespiti ve konuşma örtüşmesi ile söz kesme gibi olguların açımlaması olarak geniş anlamda tutarsak, konuşmalarda sık sık süreksiz yapılara rastlamak mümkündür (örn. *Nuray hanım hani geçende geldi telefon etti biraz önce*). STD'de bu yapılar ana birimin parçası olarak ele alınmıştır; ancak sözcelerin tümce yapısı açımlamaları yapıldığında bunların işaretlenmesi ayrı bir araştırma konusu olacaktır. Aynı şekilde sözcük türü işaretleme çalışmaları da ileriki aşamalarda ele alınacaktır. Söylem düzlemini ilgilendiren sözeylem açımlamaları ise hizmet alımlarını kapsayacak biçimde yapılmıştır.

Yazılı metinlerde açımlama konusunda ise öncelikli sorun, varolan işaretleyicilerin kapsamlı derlemlerde kullanılabilir olmayışıdır. Farklı alanlarda, çok sayıda metin içeren TUD gibi derlemler üzerinde sınanmadıkça hiçbir işaretleyici yazılımın genel kullanıma uygun olduğunu söyleyemiyoruz. Yine, alanyazındaki belirginleştirme çalışmaları da her tür metinde, benzer ya da tekrarlanabilir sonuçlar üretmediğinden, deneysel dilbilim çalışmalarından çok, bilgisayar mühendisliği uygulamaları olarak nitelendirilmeleri daha doğru olacaktır. Bu konuda yapılacak en doğru şey, çekimli/türemiş ve kök sözcükleri ve açımlamalarını içeren Türkçe'nin doğal dil işleme (DDİ) sözlüğünün oluşturulması ve etkileşimli bir veritabanı biçiminde güncellenmesi olacaktır. Böylesi bir hedef, sözcükbirim belirleme, açımlama, belirginleştirme gibi konularda Türkçe'nin de kendisine özgü ölçünleştirmelerini belirlemesini sağlayacaktır.

Örneğin sözcükbirim (İng. token) belirleme konusunda, genel amaçlı uygulamalar çoğunlukla <Mersin'in> gibi bir sözcüğü İngilizce <I've> gibi değerlendirerek iki ayrı sözcükbirime ayrıştıracaktır. Açımlama konusunda ise; hangi eklerin ayrıştırılacağı ve gösterimlerinin ne olacağı, açımlanmış metinlerin çıktı biçimi, sözcük türü belirleme gibi konularda Türkçe için ortak bir yönelim henüz yoktur. Bütün bu alandaki konular, güncel dilbilim alanyazınına değil, daha çok kullanılan yazılımın gereklerine göre şekillenmektedir (örn. Akın ve Akın, 2007; Çöltekin, 2010; Mersinli ve Aksan, 2011). Belirginleştirme çalışmaları açısından bakıldığında ise, kolaylıkla şunu söyleyebiliriz: başarım oranlarının tekrar edilebilirliği tartışmalıdır. Yine, Türkçe'de yaklaşık %50 gibi verilen belirsizlik oranı, kullanımda olmayan belirsizlikleri de içermektedir. Örneğin; herhangi bir yazılım "yüksek, yelken" gibi sözcükleri "yük+sek" ya da "yel+(i)ken" gibi iyi ayrı biçimbirime ayırabilir ancak kullanımda bu seçeneğe rastlamak, kapsamlı bir derlemde sınandığında, pek mümkün değildir. Uzun başsözcüklerin, daha öte çözümlemeleri engellediği gibi bir genellemeye de gidilemez çünkü "anı" gibi sözcüklerin "anı" ve "an+ı" biçimindeki her iki kullanımına da rastlanır ve sözcük düzeyinde belirsizlik korunur. Bu konuda yapılması gereken, Türkçe'nin sözcük düzeyinde kullanımdaki belirsizliklerinin saptanmasıdır. Bu aşamadan sonra yapılacak, eğitici derleme dayalı ya da istatistik temelli belirginleştirmelerin sonuçları tekrarlanabilir olacaktır.

Genel olarak belirtmek gerekirse, Türkçe'deki DDİ çalışmaları nicel olarak fazla gibi görünse de, standartlaşma, kullanılabilirlik ve başarım konularında yetersizdir. DDİ'nin her alt-alanında, Türkçe için çalışmalar yapılmış gibi görünse de, bunların, uygulamada ya da dilbilim çalışmalarında kullanıldığını söyleyemiyoruz ne yazık ki. Nitelik ve kullanılabilirlik açısından yeterli çalışmalara, deneysel (İng. empirical) dilbilimde duyulan gereksinim henüz giderilmiş değildir.

TUD ekibinin devam etmekte olan çalışmalarından biri de BNCweb arayüzünü kendisine örnek alan, IMS Corpus Workbench (CWB) (Evert ve Hardie, 2011) ve MySQL ilişkisel veritabanı sorgulama teknolojisini kullanan CQPweb (Hardie, 2012) derlem işleme sistemine Türkçe için oluşturulmuş derlemlerin, derlemi kuran kişi ya da ekibin izni alınarak aktarılmasını sağlamaktır (bkz. Resim 2). Bu yolla çeşitli amaçlar için hazırlanmış ancak sınırlı sayıda araştırmacının kullandığı derlemler daha çok kullanıcının sorgu yapabileceği bir platformda yer alabilecektir.

CQPweb kurulduğu sistemin özelliklerine bağlı olarak yaklaşık 2 milyar sözcüklük derlemleri işlemleyebilmektedir. Ancak sözcük sayısı arttıkça işlemleme süresi de doğru orantılı olarak artmaktadır. Tek bir derlemin gösterimini sağlayan BNCweb arayüzünden farklı olarak, CQPweb farklı derlemlerin yüklenebildiği, çeşitli işaretlemeler yapılan metinlerin (örn. biçimbilimsel-anlambilimsel) yapılan işaretlemeler bağlamında incelenmesine olanak tanımaktadır. Bu özelliğine ek olarak yapılacak tümce sonu/paragraf işaretlemelerine bağlı olarak araştırmacılar tümce/paragraf temelinde anahtar sözcük/sözcük grubu aramalarını CQPweb kullanarak gerçekleştirebilirler. Yerel bilgisayar üstünde çalışmasının yanında sunucu-istemci iletişim modelini kullanarak, arayüzü sayesinde internet üzerinden de araştırma yapmaya olanak sağlamaktadır (bkz. http://www.turkishcorpora.com).

Resim 2. Türkçe derlemler için CQPweb Arayüzü

CQPweb temel olarak hem dil araştırmacılarının hem de sistem yöneticilerinin gereksinimlerine cevap verecek bir biçimde tasarlanmıştır. Bağımlı dizin sorgulama, eşdizimlilik listeleri, dağılım listeleri, sıklık listeleri ve grafikleri oluşturma, sorgu sonuçlarının rastlantısal olarak belirli bir sıklık sayısında ya da yüzdelik oranda araştırmacılara sunulması, sorgu sonuçlarının sayısal verilerinin belirlenen metinsel ölçütlere bağlı olarak dağılımını göstermesi, çeşitli dil dışı ölçütlere bağlı olarak (örn. Yayın yılı, metin türü) yapılan sınıflandırmalara bağlı aramalar gerçekleştirilmesine ve sorgu sözcüğünün sağında ve solunda yer alan 5 sözcüğe göre alfabetik listelemeye izin vermesi dilbilim araştırmalarında kullanabilecek araçlardır. Diğer yandan, TEI standartlarına uygun olarak hazırlanmış derlemlerle uyumlu çalışma, metinler uygulamanın istediği formata aktarıldıktan sonra kolay derlem kurulumuna izin verme, kullanıcı yönetimini kolaylaştırma, Unix ve MacOs X işletim sistemleri üstünde sunucu-istemci modelini sağlayarak erişim sağlama ve gelişmiş yazılım bilgisine sahip kişilere mevcut yapıyı düzenleyebilme sistem yöneticisi ve yazılım mühendislerinin işlerini kolaylaştırmaktadır.

Derlemler CQPweb arayüzüne aktarılırken Türkçe için henüz derlem oluşturma, işleme, yayınlama ve kullanım ilkelerinin, veribilgisi ve açımlama için ontolojilerin yeterince gelişmemiş olmasının ve Türkçe derlemlerde işaretleme ve açımlama çalışmalarına gereken önemin verilmemesinin belli başlı güçlükler olduğunu rahatlıkla söyleyebiliriz.

4. Sonsöz

Bu yazıda Yuvarlak Masada tartışılan konuların sadece bir kısmı üzerinde durabildik. Yuvarlak masa tartışmaları ile amacımız Giriş bölümünde dile getirdiğimiz gibi, derlem araştırmaları yöntemleri ve çalışma uygulamaları üzerine ön plana çıkan bazı konuları ele almaktı. Bu amacımızı günümüzde sıkça kullanılan bir ifade ile özetlersek, Türkçe ve Türkiye'de yapılan derlem çalışmalarında "en iyi uygulamaları" geliştirmeye gereksinim vardır. Umuyoruz bu uygulamaların ne olduğunu tartışmaya devam ederiz.

Teşekkür

STD, TÜBİTAK 108K283 (2008-2010) ve ODTÜ, BAP-05-03-2011-001 (2011-2013) tarafından desteklenmiştir. TUD, TÜBİTAK 108K242 (2008-2011), ME.Ü, BAP- FEF İDEB (SYA) 2009-3 (2009 - 2010), ME.Ü., BAP-FEF İDEB (MA) 2009-3 (2009-2010) ve ME.Ü., BAP-FEF İDEB (SYA) 2010-3 (2010 -2011) tarafından desteklenmiştir.

Kaynakça

Adolphs, S., Knight, D. (2010). Building a spoken corpus: What are the basics? A. O'Keefe ve M. McCarthy, (Ed.), *The Routledge handbook of corpus linguistics* içinde (ss. 38-52). London/New York: Routledge.

Akın, M. D. ve Akın, A. A. (2007). Türk Dilleri için açık kaynaklı Doğal Dil İşleme kütüphanesi: ZEMBEREK. *Elektrik Mühendisliği*, 431, 38.

Aksan, Y., Aksan, M., Koltuksuz, A. ve diğerleri (2012). Construction of the Turkish National Corpus (TNC). *Proceedings of Eight International Conference on Language Resources and Evaluation* (LREC2012). 25 Ekim 2012 tarihinde http://www.lrecconf.org/proceedings/lrec2012/papers.html adresinden erişildi.

Aksan, Y. ve Demirhan, U. U. (bu kitap). Türkçe Ulusal Derlemi arayüz özellikleri: Tanıtım ve uygulama.

Anthony, L. (2012). AntConc (Version 3.3.5w) [Yazılım]. Tokyo: Waseda University.

Baroni, M. ve Evert, S. (2009). Statistical methods for corup exploitation. A. Lüdeling, M. Kytö (Ed.), *Corpus linguistics: An international handbook, vol. II* içinde (ss. 777-802). Berlin/New York: Walter de Gruyter.

Biber, D., Conrad, S. ve Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Jones, J. K. (2009). Quantitative methods in corpus linguistics. A. Lüdeling, M. Kytö (Ed.), *Corpus linguistics: An international handbook, vol. II* içinde (ss. 1287-1304). Berlin/New York: Walter de Gruyter.

Cassidy, S. (2013). Interoperable annotation in the Australian National Corpus. *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation isa-9* (ss. 35-40). Potsdam. 15 Eylül 2013 tarihinde http://sigsem.uvt.nl/isa9/ISA-9_proceedings.pdf adresinden erişildi.

Cassidy, S., Haugh, M. Peters, P. ve Fallu, M. (2012). The Australian National Corpus: National Infrastructure for Language Resources. *Proceedings of Eight International Conference on Language Resources and Evaluation* (LREC2012). 15 Eylül 2013 tarihinde http://lrec.elra.info/proceedings/lrec2012/pdf/400_Paper.pdf adresinden erişildi.

Çelebi, H. (2012). *Extracting and analyzing impoliteness in corpora a study based on the British National Corpus and the Spoken Turkish Corpus*. Doktora tezi, ODTÜ, Ankara.

Çöltekin, Ç. (2010). A freely available morphological analyzer for Turkish, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010).* 20 Ekim 2011 tarihinde http://www.lrec-conf.org/proceedings/lrec2010/bibtex.html adresinden erişildi.

Doğan, M. U. ve diğerleri (2011). *Ulusal konuşma ve dil teknolojileri platformu kuruluşu: Türkçede mevcut durum çalıştayı bildirileri*. Gebze: TÜBİTAK-TÜSSİDE, TÜBİTAK-BİLGEM, Multisaund.

Gries, S. Th. (2013). Sources of variability relevant to the cognitive sociolinguist, and corpus- as well as psycholinguistic methods and notions to handle them. *Journal of Pragmatics*, 52, 5-16.

Gries, S. Th. (2009). *Quantitative corpus linguistics with R: A practical introduction*. New York/London: Routledge.

Evert, S. ve Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. *Proceedings of corpus linguistics 2011*. 20 Kasım 2012 tarihinde http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2011-birmingham.aspx adresinden erişildi.

Hardie, A. (2012). CQPweb- combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17, 3, 308-409.

Hedeland, H. ve Wörner, K. (2012). Experiences and problems creating a CMDI profile from an existing metadata schema (ss. 37-40). Proceedings of Eight International Conference on Language Resources and Evaluation (LREC2012). 15 Eylül 2013 tarihinde http://www.lrecconf.org/proceedings/lrec2012/papers.html adresinden erişildi.

Kilgarriff, A. ve Grefenstette, G. (2003). Web as corpus: Introduction. *Computational Linguistics*, 29, 3: 333-347.
Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. M. Hundt, N. Nesselhauf ve C. Biewer (Ed.), *Corpus linguistics and the web* içinde (ss. 133-149). Amsterdam/New York: Rodopi.
Leech, G. (2011). Principles and applications of corpus linguistics. V. Viana, S. Zyngier ve G. Barnbrook (Ed.), *Perspectives on corpus linguistics* içinde (ss. 155-170). Amsterdam/Philadelphia: John Benjamins.
McEnery, T., Hardie, A. (2012). *Corpus linguistics: method, theory and practice*. Cambridge: Cambridge Unversity Press.
McEnery, T., Xiao, R. ve Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.
Mersinli, Ü. ve Aksan, M. (2011). Türkçenin biçimbirim ve sözcük türü işaretlemesi. Ç. Hatipoğlu ve Ç. Sağın-Şimşek (Ed.), *24. ulusal dilbilim kurultayı bildiri kitabı* (s. 367-376) Ankara: ODTÜ Yayınları.
Rehm, G., Witt, A. (2007). Digital text resources for the humanities – Legal issues. *Proceedings of Digital Humanities 2007*. 30 Ağustos 2013 tarihinde http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.129.2070&rep=rep1&type=pdf adresinden erişildi.
Ruhi, Ş., Eröz-Tuğa, B., Hatipoğlu, Ç., Işık-Güler, H., Acar, M. G. C., Eryılmaz, K. ve diğerleri (2010). Sustaining a corpus for spoken Turkish discourse: Accessibility and corpus management issues. *Proceedings of the LREC 2010 Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management* içinde (ss. 44-47). Paris: ELRA. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf#_page=52
Ruhi, Ş. ve Işık-Taş, E. (basımda). Constructing general and dialectal corpora for language variation research: Two case studies from Turkish. T. Schmidt, K. Wörner, Ş. Ruhi ve M. Haugh (Ed.), *Best practices in spoken corpora in linguistic research*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Ruhi, Ş. (bu kitap). Sözlü Türkçe Derlemi'nde temel arama ve edimbilimsel açımlama: Yöntem geliştirme.
Say, B., Zeyrek, D., Oflazer, K. ve Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. K. İmer ve G. Doğan (Ed.), *Current research in Turkish linguistics: proceedings of the 11th International Conference of Turkish Linguistics* içinde (ss. 183-192). Magusa: Eastern Mediterranean University.
Silberztein, M. (2003). *User manual*.26 Eylül 2013 tarihinde http://www.nooj4nlp.net/ adresinden erişildi.
Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*, 1. http://jtei.revues.org/142, doi: 10.4000/jtei.142
Schmidt, T. (basımda). (More) common ground for processing spoken language corpora? T. Schmidt, K. Wörner, Ş. Ruhi ve M. Haugh (Ed.), *Best practices in spoken corpora in linguistic research*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Sinclair, J. McH. (1994). *Trust the text: Language, corpus and discourse*. London: Routledge.
Sinclair, J. McH. (2005). How to build a corpus. M. Wynee (Ed.), *Developing linguistic corpora: A guide to good practice* içinde (ss. 96-101). 30 Nisan 2013 tarihinde http://www.ahds.ac.uk/guides/linguistic-corpora/appendix adresinden erişildi.
Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins.
Válková, L., Waclawičová, M.ve Křen, M. (2012). Balanced data repository of spontaneous spoken Czech. N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani ve diğerleri (Ed.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC) (ss. 3345-3349). 12 Ocak 2013 tarihinde http://www.lrec-conf.org/proceedings/lrec2012/pdf/179_Paper.pdf adresinden erişildi.