Mersin Üniversitesi Dil ve Edebiyat Dergisi, MEUDED, 2017; 14 (2), 1-14.

GÜNCEL BİÇİMİYLE SÖZLÜ TÜRKÇE DERLEMİ: TEKNİK VE İSTATİSTİKSEL BİR ÇÖZÜMLEME

Güneş Acar[1]

KU Leuven

Öz: Bu makalenin öncelikli amacı Prof. Dr. Şükriye Ruhi'nin öncülüğünde geliştirilen ODTÜ Sözlü Türkçe Derlemi'nin (STD) oluşturulmasında kullanılan teknolojileri ve iş akışlarını açıklamaktır. STD'nin oluşturulmasında merkezi bir konumda olan Web Tabanlı Derlem Yönetim Sistemi, kayıtların çeviriyazısı, kontrolü ve yayınlanmasını kolaylaştıran bir dizi iş akışını, veri biçimini ve dışa aktarma seçeneklerini barındırmaktadır. Derlem yönetim sistemi, proje araştırmacıları tarafından Python programlama dili kullanılarak geliştirilmiş olup, farklı rollere sahip proje üyelerinin çevrimiçi bir arayüzden uzaktan ortaklaşa çalışabilmelerini sağlamaktadır. STD kapsamında 286,391 sözcüklük konuşmanın çeviriyazısı tamamlanmış ve kontrolden geçmiş, 79,189 sözcüklük konuşma ise bütünüyle yayına hazır biçime getirilmiştir. Makalede derlemdeki bu kayıtlarla ilgili genel istatistikler sunularak STD'nin daha geniş çaplı bir sürümü için yapılması gerekenler tartışılmaktadır.

Anahtar sözcükler: *Sözlü derlem, derlem yönetim sistemi, EXMARaLDA*

[1] KU Leuven, Elektronik Mühendisliği Bölümü, Leuven, Belçika, gunes.acar@esat.kuleuven.be
Makale gönderim tarihi: 27 Haziran 2017; Kabul tarihi: 26 Kasım 2017

SPOKEN TURKISH CORPUS IN ITS PRESENT FORM: A TECHNICAL AND STATISTICAL ANALYSIS

Abstract: The primary goal of this article is to explain the technologies and workflows used to build the METU Spoken Turkish Corpus (STC), which is pioneered by the late Prof. Dr. Şükriye Ruhi. The Web Based Corpus Management System, which is crucial to the building of STC, contains a set of workflows, data formats and export options that make it easy to transcribe, control and publish corpus data. Corpus Management System was developed by the STC project members using the Python programming language and it enables the collaboration of remote project members with different roles through an online interface. Within the STC, 286,391 words long speech are transcribed and checked; in addition, 79,189 words long recordings are made ready to publish. The article presents general statistics about the recordings in the STC and discusses what needs to be done for the publication of a large scale version of the STC.

Key words: *Spoken corpus, corpus management system, EXMARaLDA*

1. GİRİŞ

Bu makale Türkçenin ilk sözlü derlemi olan ODTÜ Sözlü Türkçe Derlemi'nin geliştirilmesinde ve yönetilmesinde kullanılan teknolojileri, yazılımları ve iş akışlarını ana hatlarıyla açıklamaktadır. Makale öncelikle STD için geliştirilen ve derlem verilerinin toplanması ve yönetilmesi açısından merkezi olan Web tabanlı derlem yönetim sisteminin tasarımında izlenen hedefleri sunmaktadır. Daha sonra güncel biçimiyle derlem yönetim sisteminin özellikleri ve kullanılan teknolojiler açıklanmaktadır.

Makalenin ikinci kısmında STD'deki kayıtlar ve çeviriyazılarla ilgili istatistikler verilmektedir. Bu kısımda ODTÜ STD'nin daha geniş çaplı bir sürümünün yayınlanabilmesi için izlenmesi önerilen bir yaklaşım sunulmaktadır. Böyle bir yayınlama girişimi için hazır biçime getirilmesi önerilen kayıtlar hakkında bilgiler ve istatistikler verilmektedir.

2. STD WEB TABANLI DERLEM YÖNETİM SİSTEMİ

STD'nin oluşturulmasında merkezi bir konumda olan web tabanlı derlem yönetim sistemi, kayıtların çeviriyazısı, kontrolü ve yayınlanmasını kolaylaştıran bir dizi iş akışını, veri biçimini ve dışa aktarma seçeneklerini barındırmaktadır (Acar ve Eryılmaz, 2010). Derlem yönetim sistemi, proje araştırmacıları tarafından Python programlama dili kullanılarak geliştirilmiş olup, farklı rollere sahip proje üyelerinin çevrimiçi bir arayüzden uzaktan ortaklaşa çalışabilmelerini sağlamaktadır. Derlemin oluşturulmasında kullanılan iş akışlarının ve kontrollerin yazılım tarafından otomatikleştirilmesi kullanıcı hatasını en aza indirgemeyi hedeflemektedir.

STD gönüllüleri tarafından toplanan ses veya video kayıtlarının ilgili kayıt izin formları ve üst veriler (İng. *metadata*) ile derlem yönetim sistemine yüklenmesini takiben kayıtlar hakkındaki tüm işlemler ve üst veriler merkezi bir ilişkisel MySQL veritabanında saklanır. Sistemde tanımlı iş akışı sayesinde derlem kayıtlarının hangi işlemlerden geçtiği, proje çalışanlarının hangi kayıtlar üzerinde çalıştığı ve iş yükleri izlenebilir.

Derlem yönetim sistemi EXMARaLDA (Schmidt, 2004) yazılım kütüphanelerinin yardımıyla, STD kayıtlarının ELAN, TEI ve PRAAT gibi derlembilimcilerin yaygın olarak kullandığı program ve biçimlerle uyumlu olarak dışa aktarabilir. Bu şekilde farklı araç ve platformlarla çalışan araştırmacıların STD'yi kullanabilmeleri olanaklıdır.

2.1. STD WEB TABANLI DERLEM YÖNETİM SİSTEMİ TASARIM HEDEFLERİ

STD Web Tabanlı Derlem Yönetim Sistemi beş ana tasarım hedefi doğrultusunda geliştirilmiştir (Acar ve Eryılmaz, 2010):

i. Çevrimiçi Erişim ve Merkezi Veritabanı: Çevrimiçi erişim, derlemdeki ses ve görüntü kayıtları, çeviriyazı ve üst verilere proje ekibi ve gönüllüler tarafından kolay erişimi sağlar. Çevrimiçi erişim, konuşma kaydı gibi hassas verilerin İnternet üzerinden gerçekleştirilebilecek herhangi bir saldırıya karşı gerekli güvenlik önlemlerinin alınmasını gerektirir. Merkezi veritabanı ise birçok kişinin aynı anda derlem verileri üzerinde eşzamanlamaya gerek kalmadan çalışmasını sağlar.

ii. İş Akışı Denetimi ve Kullanıcı Rolleri: Derleme eklenen her kayıt Tablo 1'de listelenen sekiz iş akışı aşamasından geçerek yayına hazır biçime gelmektedir. Derlem yönetim sistemi bu aşamaların doğru sırayla, doğru kişiler tarafından ve gerekli kontroller yapılarak geçildiğini garanti eder. Bu sayede gönüllülerin sisteme eklenmesi, kayıtların toplanması, sınıflandırılması ve kontrolü, çeviriyazı sürecinin ve çıktılarının kontrol edilmesi gibi aşamalar, proje ekibi tarafından kolaylıkla çevrimiçi arayüz üzerinden takip edilebilir.

STD'de her proje üyesi bir ya da daha fazla role sahiptir. Bu roller kullanıcıların sistemdeki yetkilerini ve iş akışı aşamalarındaki sorumluluklarını belirler. Örneğin çeviriyazı kontrolü sadece *değerlendirici* (İng. *evaluator*) rolüne sahip proje üyeleri tarafından yapılabilirken, kayıtların çeviriyazılara atanması görev atayıcısı (İng. *task asssigner*) üyeler tarafından yapılır. Her bir iş akışı aşamasının sonunda ilgili dosya çevrimiçi arayüzden yüklendiğinde sıradaki aşamaya bağlı olarak ilgili role (Örn. çeviriyazı kontrolü) sahip kullanıcıya otomatik bir e-posta gönderilir. Bu sayede zaman geçirmeden ve aşamalar atlanmadan bir sonraki aşamaya geçilmesi garanti edilmiş olur.

Tablo 1. ODTÜ STD'de kullanılan iş akış aşamaları

Aşama adı
Çeviriyazıcı atanmadı (Not assigned)
Çeviriyazıcı atandı (Assigned)
Sürüyor (In progress)
Çeviriyazı kontrol bekliyor (Waiting for transcription feedback)
Düzeltme (In correction)
Son kontrol için bekliyor (Waiting for final check)
Tamamlandı (Completed)
Yayınlanabilir (Publishable)

iii. Saydamlık ve Tutarlılık: STD Web Tabanlı Derlem Yönetim Sistemi'nin tasarımında, iş akışı ve yönetim süreçlerinin açık bir şekilde takip edilebilmesi ve kendi içinde tutarlı olması hedeflenmiştir. Kullanıcı bilgileri ve etkinlikleri, çeviriyazı istatistikleri gibi değişkenler her an ulaşılabilir biçimdedir.

Örneğin, yeni bir kaydın çeviriyazıcılara atanması aşamasında her çeviriyazıcının üzerindeki iş yükü sistemde sorgulanarak kaydın en

uygun kişiye atanması sağlanır. Aynı biçimde yaptıkları çeviriyazı saati üzerinden ücret alan çeviriyazıcıların kaç saat çeviriyazı yaptığı listelenerek ücretleri buna göre ödenebilir.

Derlem oluşturulmasında gözetilen temel özelliklerden olan dengelilik (İng. *balancedness*), derlemdeki metin türlerinin gerçek dil evrenindeki kullanımları oranında temsil edilmesi olarak açıklanmıştır (Leech, 2007). Derlemdeki kayıtların dengeli dağılımı da yine çevrimiçi arayüzden erişilebilen istatistikler yardımıyla sağlanır. Bu arayüz üzerinden kayıtlar ve çeviriyazıların konuşma ortamı, konuşma türü, kaydın yapıldığı coğrafi bölge gibi birçok değişkene bağlı dağılımları anlık olarak erişilebilir. Bu anlık dağılımlar derlemin hedefleriyle (Ruhi ve diğ., 2010) karşılaştırılarak ne tür kayıtların toplanması gerektiği planlanabilir.

iv. Genişletilebilirlik: Derlem yönetim sistemi, EXMARaLDA sayesinde sistemdeki çeviriyazıları birçok veri formatında dışa aktarabilir. Çeviriyazılar ve üst veriler dahili olarak MySQL veri tabanında saklandığından yeni üst veri türleri ve dosya biçimleri için destek kolayca eklenebilir. Böylece derlemin farklı platformlarda kullanılabilirliği arttırılmış olur.

v. Uzun Vadede Kullanılabilirlik: Derlemin uzun vadede ne kadar kullanılabilir olduğu derlemin sunumunda kullanılan veri biçimleri ve standartlarla bağlantılıdır. Derlemi yayınlarken kullanılan HTML ve XML standartları, daha genel ve ortadan kalkması ihtimali düşük dosya tipleri olarak sistemin uzun vadede kullanılabilirliğini garanti eder.

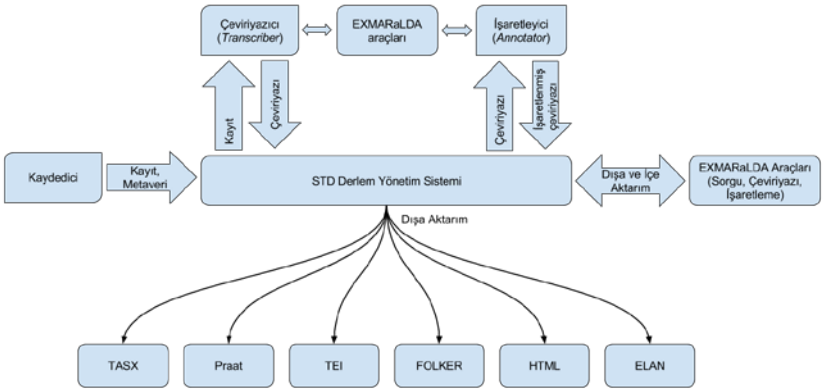
2.2. STD WEB TABANLI DERLEM YÖNETİM SİSTEMİ ÖZELLİKLERİ

STD Web Tabanlı Derlem yönetim sistemini geliştirirken farklı birçok yazılım ve enformasyon teknolojisinden faydalanılmıştır. Derlem yönetim sistemi'nin çekirdeğini Web2py adlı Python tabanlı bir web uygulama iskeleti (İng. *framework*) oluşturmaktadır (DiPierro, 2009). Web2py, kimlik doğrulama, hata ayıklama, veritabanı yönetimi gibi hazır modülleri ve MVC mimarisi desteği sayesinde web uygulamalarının güvenli ve hızlı bir biçimde geliştirilmesine olanak tanır (DiPierro, 2011).

Kayıt ve çeviriyazı dosyaları Subversion (SVN) sürüm yönetim sistemi ile otomatik olarak sürüm kontrolü altına alınır. Bu sayede dosyadaki değişiklikler ve bu değişikliklerin hangi an, kim tarafından yapıldığı kayıt altına alınır ve gerekirse önceki sürümlere dönülebilir.

EXMARaLDA araçları entegrasyonu sayesinde çeviriyazılar TEI, Praat, ELAN, TASX Annotator, HTML, PDF, RTF biçimlerinde dışa aktarılabilir (bkz. Şekil 1). Bu sayede STD verileri birçok başka yazılım kullanılarak çözümlenebilir.

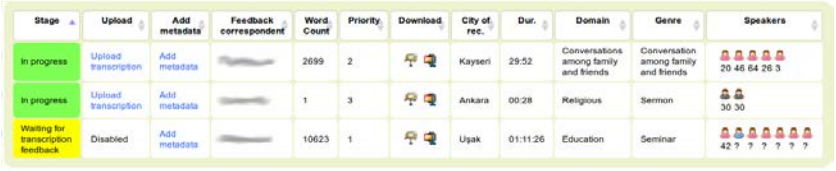
Derlemdeki veriler, toplam 118 tablodan oluşan bir MySQL ilişkisel veritabanında saklanmaktadır. Bu tabloların 54'ü ODTÜ STD web sitesinde kullanılan Wordpress yazılımı tarafından kullanılmaktadır. Derlemdeki çeviriyazılar, kullanıcılar, iş akışları, etkinlikler ve bunlarla ilgili üst veriler geri kalan 64 tabloda saklanır. Bunlara ek olarak derlem verilerinin ikinci bir kopyası test veritabanında tutulmaktadır. Derlem yönetim sistemi yazılımında yapılan değişiklikler öncelikle test veri tabanında denenir ve olası yazılım hatalarının derlem verilerine zarar vermesi engellenmiş olur.



Şekil 1. STD iş akışı ve popüler derlem yazılımları ve dosya biçimleri ile birlikte çalışabilirlik

Çevrimiçi arayüzü, proje üyelerinin, çeviriyazıcıların ve gönüllülerin aynı anda uzaktan çalışmasına olanak verirken insan hatasını en aza indirgemek amacıyla tüm iş akışlarını otomatize etmeye çalışır. Aynı zamanda kullanıcıların üzerine çalıştıkları kayıt ile ilgili bilgiye hızlıca erişmesi sağlanmıştır. Örneğin, Şekil 2'de gösterilen çeviriyazılarındaki

kayıt üst verileri, konuşmacıların yaşları, sayısı ve cinsiyetleri kullanılan ikonlar sayesinde kolayca anlaşılabilmektedir.



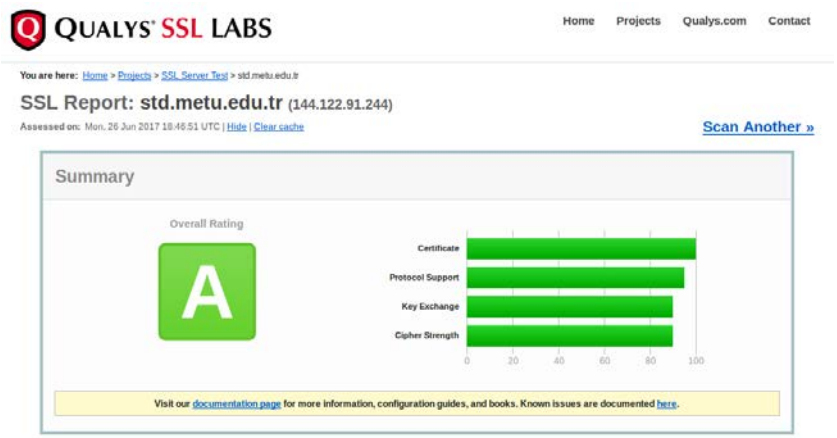
Şekil 2. Çeviriyazı listesi arayüzü

Web tabanlı sistemin bir başka avantajı Windows, MacOSX, Linux gibi farklı işletim sistemleri kullanan proje çalışanlarının herhangi bir ek program kurmadan yalnızca web tarayıcısı yardımıyla sisteme erişebilmesidir.

*Güvenlik***:** Sistemde İnternet üzerinden gelebilecek saldırılara karşı birkaç düzeyde önlem alınmıştır. İşletim sistemi düzeyinde kararlılığı ile bilinen Debian Linux dağıtımı kullanılmaktadır. Sistem, güvenlik yamaları ve yazılım güncellemeleri yüklenerek sürekli güncel tutulmaktadır. Web sitesinin yapıldığı Wordpress yazılımı ise güvenlik amaçlı WordFence[2] eklentisi ile korunmaktadır. WordFence saldırıları tespit ederek, saldırganların IP adreslerini engellerken, saldırganların IP adreslerini günlük olarak yönetici e-posta adresine rapor eder. Aynı yazılım, güncellenmesi gereken Wordpress eklentilerini anında rapor ederek web sitesinin güvenli ve güncel kalmasına yardımcı olur.

Gizlilik: ODTÜ STD'de saklanan veriler konuşmacıların ses kaydını içerdiğinden bu kayıtların ve ilgili üst verilerin gizliliği çok önemlidir. Şifrelemenin kullanılmaması durumunda çeviriyazı dosyaların indirilmesi sırasında ağdaki paketleri izleyen bir saldırganın kayıtları veya kayıtlara erişimi olan kullanıcıların şifrelerini ele geçirmesi olasıdır. Bunu engellemek için ODTÜ STD sitesinin tüm sayfalarında HTTPS bağlantısı kullanır. ODTÜ STD, web sitelerinin SSL bağlantı güvenliğini çevrimiçi olarak test eden Qualys SSL Labs'dan A notu almıştır (bkz. Şekil 3).

[2] https://www.wordfence.com/

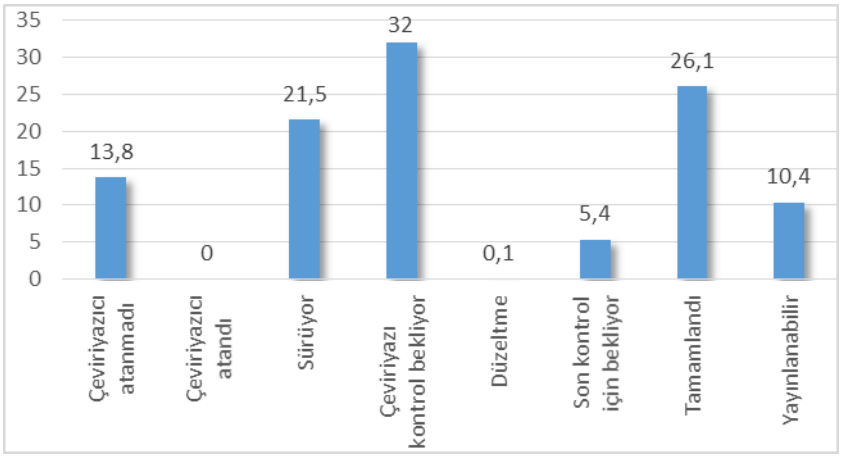


Şekil 3. Qualys SSL Labs'ın ODTÜ STD web sitesi SSL raporu.

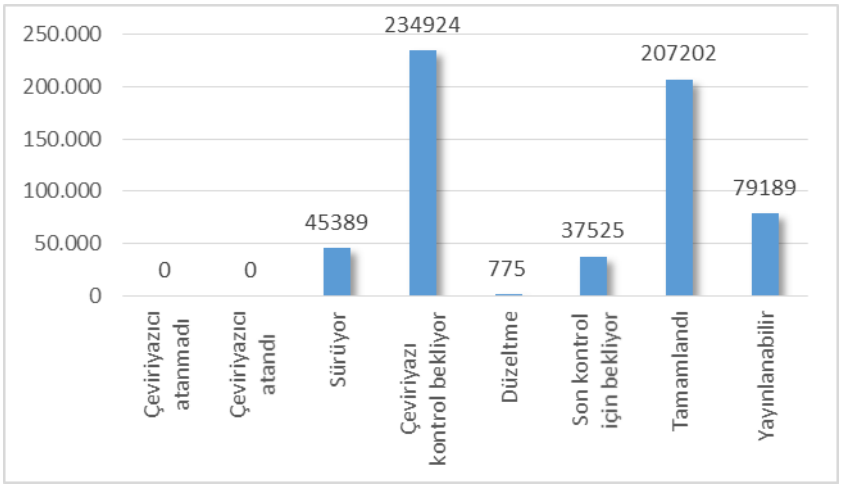
3. SAYILARLA SÖZLÜ TÜRKÇE DERLEMİ

STD kapsamında 286,391 (36.5 saat) sözcüklük konuşmanın çeviriyazısı tamamlanmış ve kontrolden geçmiş, 79,189 sözcüklük (10.4 saat) konuşma ise bütünüyle yayına hazır biçime getirilmiştir. Çeviriyazısı devam eden veya yarım kalan kayıtlar da sayıldığında STD'de çeviriyazısı yapılmış 605,004 sözcüklük kayıt bulunmaktadır. Derlemde her bir kayıt için bölge, konuşanların yaşları ve cinsiyetleri, meslekleri, nerede yaşadıkları, kaydın nasıl bir ortamda alındığı gibi birçok üst veri tutulmaktadır. Aşağıda STD'nin bu üst verilere bağlı dağılımına dair istatistikler sunulmuştur.

Kayıtların Çeviriyazı Aşamalarına Göre Dağılımı: STD'deki kayıtların çeviriyazı aşamalarına göre dağılımı Şekil 4'te verilmiştir. Buna göre 32 saatlik kayıt kontrol için beklemektedir. Çeviriyazıları tamamlanan kayıtların toplam uzunluğu ise 26.1 saatken, 10.4 saatlik kayıt tüm aşamalardan geçmiş ve yayına hazır olduğu onaylanmıştır. STD'nin gelecek sürümlerinde yayınlanması diğer aşamalardaki kayıtlara göre daha kolay olan tamamlanmış ve yayına hazır kayıtlara odaklanılabilir.



Şekil 4. Kayıt Sürelerinin Çeviriyazı Aşamalarına Göre Dağılımı (Saat olarak)



Şekil 5. Kayıtların Sözcük Sayılarının Çeviriyazı Aşamalarına Göre Dağılımı

Şekil 5 STD'deki kayıtların farklı çeviriyazı aşamalarına göre dağılımını bu defa sözcük sayısı olarak göstermektedir. Buna göre toplam 605,004 sözcüklük kaydın 207,202 sözcüklük kısmının çeviriyazısı tamamlanmış, 79,189 sözcüklük kayıt ise tümüyle yayına hazır biçime getirilmiştir. Bu iki aşamada bulunan toplam 286,391 sözcüklük kaydın yeniden çeviriyazı yapılmasına gerek olmadan STD'nin yeni bir sürümü olarak yayınlanması olanaklı görünmektedir.

STD'deki kayıtların çok daha büyük bir kısmı (234,924 sözcük) çeviriyazı kontrolü için beklemektedir. Bu kayıtların yayına hazır biçime getirilmesi ciddi bir çeviriyazı değerlendiricisi emeği gerektirmektedir. Bu değerlendirmelerin daha öncekilerle aynı standartlarda yapılması yayınlanacak verilerin tutarlılığı ve kalitesi açısından kritiktir.

Kayıtların Konuşma Ortamlarına Göre Dağılımı: STD'deki kayıtların daha iyi temsil edebilirlik (İng. *representativeness*) için farklı konuşma ortamlarından belirlenmiş oranlarda toplanması hedeflenmiştir (Biber, 1993; Ruhi ve diğ., 2010). Kayıtların bu konuşma ortamlarına (İng. s*peech domain*) göre dağılımı aşağıda Tablo 2'de verilmiştir.

Tablo 2. Derlemdeki Kayıtların Konuşma Ortamı Dağılımı (Saat olarak)

	Aile ve/veya akraba ortamı	İş ortamı	Eğitim	Radyo/TV yayınları	Arkadaş/tanıdık ortamı	Hizmet alımı	Aile ve arkadaşlarla konuşmalar	Araştırma	Yabancılar arasında kısa süreli	Sınflandırılmamış	Dini	Yasal	Politik	Kamusal	Toplam
Kayıt	17.1	11.6	16	13.3	18.3	2.9	9.7	6.2	0.4	0	0	0	0	0	95.5
Hedef	12.5	10	7.5	7.5	6	2.5	2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	50
İlerleme (%)	136	116	213	177	305	117	487	2135	130	0	0	0	0	0	191

STD kayıtlarının neredeyse tüm ortamlar için hedefin üzerinde olmasına rağmen bu oranların çeviriyazı aşamalarından bağımsız olarak hesaplandığı unutulmamalıdır. Bir başka deyişle Tablo 2'deki bu sayılar hem yayına hazır hem de henüz çeviriyazısına başlanmamış kayıtları içermektedir.

STD'de tanımlı 14 konuşma ortamından *Arkadaş/tanıdık ortamı, Aile ve arkadaşlarla konuşmalar ve Araştırma* ortamlarından hedeflenenden çok daha fazla (sırasıyla %305, %487, %2135) kayıt yapıldığı görülmektedir. Bu ortamlardaki kayıtların oranının fazla olmasının nedeni olarak, kayıt yapan ve çoğu üniversite öğrencisi olan gönüllülerin bu ortamlardan daha kolay kayıt alması gösterilebilir. Unutulmamalıdır ki, derlemin yayını sırasında çeviriyazılar daha dengeli seçilerek kayıtlardaki dengesizliğin derleme etkisi engellenebilir.

STD'de her bir konuşma ortamı, kendi içinde farklı konuşma türlerine ayrılır. STD'de tanımlı 76 farklı konuşma türünden bazıları Tablo 3'de listelenmiştir.

Tablo 3. STD'de kullanılan bazı örnek konuşma türleri

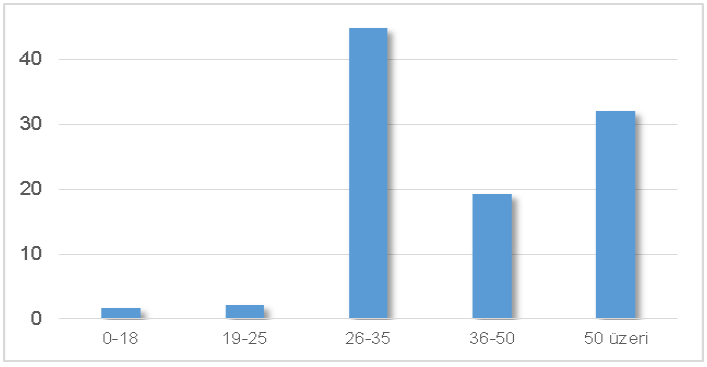
Konuşma türü	Konuşma Ortamı
Aile ve/veya akrabalar arasında sohbet	Aile ve/veya akraba
Arkadaşlarla ders çalışma	Arkadaş/tanıdık
İş toplantısı	İş
Alışveriş	Hizmet alımı
Seminer	Eğitim
Konferans	Eğitim
Siyasi hitap	Politik

Konuşma türlerine bağlı istatistiklere çevrimiçi erişim sayesinde verili bir konuşma ortamı için kayıtların konuşma türlerine göre dağılımına bakılarak konuşma türlerinin derlemde dengeli temsili hedeflenebilir.

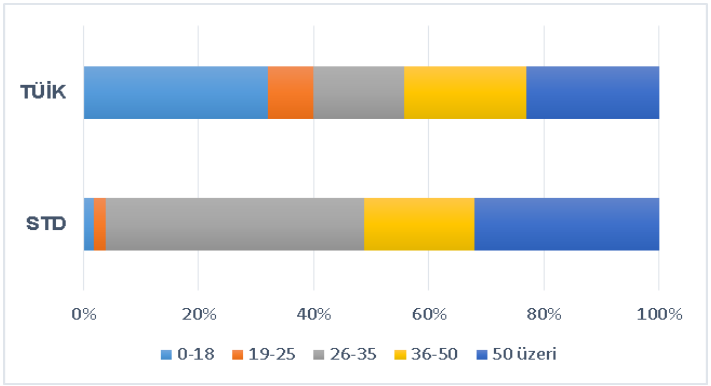
Konuşmacıların Yaş ve Cinsiyete Göre Dağılımı: STD'de toplam 686 konuşmacının kaydı bulunmaktadır. Derlem yönetim sisteminin çevrimiçi istatistikler arayüzünde konuşmacıların yaş ve cinsiyete göre dağılımları sunularak bu değişkenlere göre dengeli kayıtların toplanması mümkündür.

Şekil 6'da konuşmacıların yaş aralıklarına göre dağılımı gösterilmiştir. Buna göre 218 konuşmacının (%31.8) yaşı bilinmemektedir. Geri kalan 468 konuşmacının 210'u (%44.9) 26-35 yaş arasındadır.

Derlemdeki konuşmacıların yaş dağılımını, benzer yaş aralıklarının Türkiye İstatistik Kurumu verilerine (TÜİK, 2017) göre dağılımı ile karşılaştırdığımızda, derlemde 0-18 yaş arası nüfusun olduğundan daha az bir şekilde temsil edildiği görülmektedir (bkz. Şekil 7). Ayrıca 26-35 yaş aralığının Türkiye nüfusuna oranla daha yüksek olarak temsil edildiği görülmektedir. Bu farklılık kayıtları yapan gönüllülerin ağırlıklı olarak üniversite öğrencileri olmasından kaynaklanıyor olabilir.



Şekil 6. STD'deki konuşmacıların yaş aralıklarına göre dağılımı



Şekil 7. Türkiye nüfusunun (TÜİK, 2017) ve STD konuşmacılarının yaş aralıklarına dağılımlarına göre karşılaştırılması[3]

[3] TÜİK istatistikleri STD'den farklı yaş aralıkları kullandığından burada sunulan oranlar yaklaşıktır. STD ve TÜİK'de kullanılan yaş aralıkları için şu eşleştirme kullanılmıştır: 0-18 => 0-19, 19-25 => 20-24, 26-35 => 25-34, 36-50 => 35-49, 51+ => 50+.

Konuşmacıların cinsiyete göre dağılımına bakıldığında konuşmacıların %53.2'sinin kadın, %46.8'inin ise erkek olduğu görülmektedir (bkz. Tablo 4). Bu veriler ışığında derlemdeki konuşmacıların cinsiyete bağlı olarak dengeli dağıldığı söylenebilir.

Tablo 4. STD'deki toplam 686 konuşmacının cinsiyete göre dağılımı

Kadın	Erkek	Bilinmeyen
358	315	13

4. TARTIŞMA VE SONUÇ

Bu makalede STD'nin oluşturulmasında kullanılan teknolojiler, veri formatları ve iş akışları açıklanmış; derlemdeki yayınlanması henüz mümkün olmamış kayıt ve çeviriyazılar ile ilgili güncel istatistikler sunulmuştur.

STD'nin 2010 yılında yayınlanan demo sürümü için 9 farklı ülkeden 100'ü aşkın araştırmacı kullanım için başvuruda bulunmuştur. STD'nin demo sürümüne araştırmacıların ilgisi dikkate alındığında çeviriyazısı tamamlanmış ve yayına hazır biçimdeki toplam 286,391 sözcüklük kaydın STD'nin yeni bir sürümü olarak yayınlanarak araştırmacılarla paylaşılması projenin planlanan amacına ulaşması açısından çok önemlidir. Çeviriyazısı devam eden veya kontrol bekleyen kayıtların tamamlanarak bu sürüme dahil edilmesi mümkün olsa da yürütücüsünü kaybetmiş bir proje için bu hedefe kısa bir sürede ulaşmak olası görünmemektedir. Bu nedenle, ilk büyük çaplı sürüm için çalışmaların gerçekçi bir hedefle yapılması önemlidir. Bunu izleyen zaman içinde gerekli çeviriyazı ve değerlendirici iş gücünün bulunması durumunda derlemdeki diğer kayıtların da yayınlanması düşünülebilir.

Çeviriyazısı tamamlanmış kayıtların konuşmacı yaş grubu, konuşma ortamı ve konuşma türü gibi değişkenlere göre dengeli seçilmesi konusunda iki seçenek bulunmaktadır: yayına hazır haldeki tüm kayıtların yayınlanması veya derlemin tasarım aşamasında belirlenen hedef oranlar dikkate alınarak örneklenen dengeli bir alt-derlemin yayınlanması. İlk seçenek daha geniş çaplı bir sürümü olası kılarken, ikinci seçenek derlemin tasarım hedeflerine daha yakın olan temsil edebilirliği daha yüksek bir derlemin yayınlanmasını sağlayacaktır.

Sözlü Türkçe araştırmaları açısından önemli bir boşluğu doldurabilecek olan ODTÜ Sözlü Türkçe Derlemi'nin daha geniş çaplı bir sürümü STD projesinin hem beyni hem de en çalışkan üyesi olan sevgili Prof. Dr. Şükriye Ruhi'nin hatırasına en iyi armağan olacaktır.

TEŞEKKÜR

Değerli yorumları ve önerileri için dergi hakemlerine; karşılığı ödeneyemecek emekleri için ODTÜ Sözlü Türkçe Derlemi'nin tasarım, kayıt, çeviriyazı, işaretleme ve diğer aşamalarında katkıda bulunan proje gönüllülerine, ODTÜ Yabancı Diller Eğitimi Bölümü 2009-11 yılları yüksek lisans öğrencilerine ve diğer araştırmacılara teşekkürü borç bilirim.

KAYNAKÇA

Acar, M. G. C. & Eryılmaz, K. (2010). Sözlü Derlem İçin Web Tabanlı Yönetim Sistemi. *24. Ulusal Dilbilim Kurultayı Bildiri Kitabı*. 17-18 Mayıs 2010, 437-443.

Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4): 243-258.

DiPierro, M. (2009). *Web2py Enterprise Web Framework*. Wiley Publishing.

DiPierro, M. (2011). Web2py for scientific applications. *Computing in Science & Engineering*, *13*(2), 64-69.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. M. Hundt, N. Nesselhauf ve C. Biewer (Haz.) içinde. *Corpus Linguistics and the Web*, Amsterdam: Rodopi, pp. 133-49.

Schmidt, T. (2004). Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*.

Ruhi, Ş., Işık-Güler, H., Hatipoğlu, Ç., Eröz-Tuğa, B., & Çokal Karadaş, D. (2010). Achieving representativeness through the parameters of spoken language and discursive features: the case of the Spoken Turkish Corpus. *Language Windowing through Corpora. Visualización del lenguaje a través de corpus.* Part II. Universidade da Coruna, 789-799.

TÜİK. *İl, yaş grubu* ve *cinsiyete göre nüfus. E*rişim Adresi: *http://www.tuik.gov.tr/PreIstatistikTablo.do?istab_id=945. Erişim tarihi: 26/06/2017.*

KISALTMA LİSTESİ

HTML: Hypertext Markup Language
HTTPS: Hyper Text Transfer Protocol Secure
MVC: Model-view-controller
IP: Internet Protocol
RTF: Rich Text Format
STD: Sözlü Türkçe Derlemi
SSL: Secure Sockets Layer
TEI: Text Encoding Initiative
XML: Extensible Markup Language