Sağlık Bilimleri Türkçe Derlemi

Memduh Çağrı Demir[1], Mehmet Kamil Sulubulut[1]
ve Atilla Aral[2]

[1] Yonca Teknoloji, Ankara, Türkiye
{cagri.demir,kamil.sulubulut}@yt.com.tr
http://www.yt.com.tr
[2] Ankara Üniversitesi Tıp Fakültesi, Ankara, Türkiye
aral@medicine.ankara.edu.tr

Özet. Günümüzde veri madenciliği ve yapay öğrenme alanlarındaki gelişmeler nedeniyle verinin önemi her geçen gün artmakta ve geliştirilen yazılımlar farklı kaynaklardan alınan verilere dayalı olmaktadır. Dilbilim alanında yapılan çalışmalar sayesinde, yazılımlar doğal dilden oluşan verileri de işleyebilmektedir. Doğal dil işlemek için kullanılan yöntemlerden birisi derlem bazlı (İng. *corpus based*) doğal dil işleme yöntemleridir. Bu çalışmada, özellikle sağlık bilimleri alanında yapılacak çalışmalarda kullanılmak üzere oluşturulan bir Türkçe derlem anlatılmıştır. Derlem kelime kökü (İng. *lemma*), sözcük türü etiketleri (İng. *part-of-speech tags*) ve kelimelerin morfolojik analizi bilgilerini içermektedir. Derlem oluşturulurken çevrimiçi olarak ulaşılabilen ve sağlık bilimleri alanında yayımlanan açık erişimli akademik dergiler kullanılmıştır. Oluşturulan sağlık bilimleri derleminin kapsamı ulusal sağlık bilimleri veri tabanı ile karşılaştırılarak ölçülmüştür. Derlem akademik çalışmalarda kullanılmak üzere bu bildirinin kaynak gösterilmesi şartıyla tüm araştırmacıların kullanımına açıktır.

Anahtar Kelimeler: Türkçe derlem; sağlık bilimleri Türkçe derlemi; Türkçe dil işleme

Turkish Corpus on Health Sciences

Abstract. Recently as a result of developments in data mining and machine learning fields, data becomes more important day by day and developed softwares rely on data collected from different sources. With the help of studies conducted in linguistics, softwares are now capable of processing natural language. Corpus based methods are one of the methods of natural language processing. In this work, a Turkish corpus aimed to be used in medical researches is introduced. The created Turkish corpus consists of lemmas, part of speech tags and morphological analysis of each word. Corpus contains only publicly available academic journals published in health sciences. Coverage of corpus is measured against the

national health sciences database. Corpus is available for all researchers provided that this report is cited.

Keywords: Turkish corpus; health sciences Turkish corpus; Turkish language processing

1 Giriş

Derlem, dil bilimlerinde belirli bir amaç için yapılandırılmış kelimelerden oluşan genellikle elektronik olarak saklanan bir kelimeler bütünüdür. Genel amaçlı ve özel amaçlı olmak üzere iki kategoriye ayrılan derlemler özellikle dilbilim çalışmalarında aktif olarak kullanılmaktadırlar. Türkçe için oluşturulmuş derlemleri incelediğimiz zaman var olan çevrimiçi Türkçe derlemlerin çoğunlukla genel amaçlı olduğunu görüyoruz. [1]

Doğal dil işleme çalışmalarının her alanda olduğu gibi sağlık bilimleri alanında da gelişmesi sonucunda, sağlık bilimleri alanında güncel bir Türkçe derlem oluşturma gerekliliği ortaya çıkmıştır. Bu çalışmada sağlık bilimleri alanında yapılacak Türkçe doğal dil işleme çalışmalarında kullanılmak üzere oluşturulan bir derlem anlatılmaktadır. Oluşturulan derlemin gelecekte yapılacak Türkçe doğal dil işleme çalışmaları için bir temel oluşturacağı ve yapılacak çalışmaları kolaylaştıracağı düşünülmektedir.

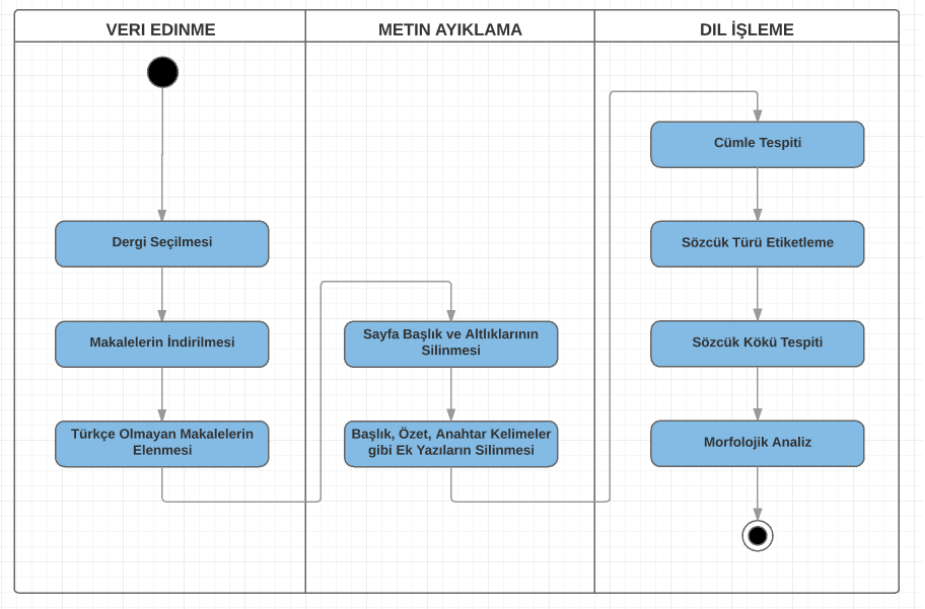
Sağlık bilimleri Türkçe derlemi oluşturulurken doğal dil işleme aracı olarak Zemberek-NLP kütüphanesinden faydalanılmıştır. Zemberek-NLP kütüphanesi, JAVA diliyle yazılmış, açık kaynak kodlu bir Türkçe doğal dil işleme kütüphanesidir. 2005 yılında 4. Linux ve Özgür Yazılım şenliğinde yılın en iyi özgür yazılımı ödülünü almıştır. Proje başlangıcında bir Türkçe imla denetim kütüphanesi olarak geliştirilmesine karşın günümüzde imla denetimine ek olarak morfolojik analiz (İng. *morphological analysis*), belirsizlik giderme (İng. *disambiguation*), dizgeciklere ayırma (İng. *tokenization*), cümle sınırları tespiti (İng. *sentence boundary detection*) ve dil tanıma (İng. *language detection*) işlemlerini yapabilmektedir. Zemberek projesi aktif olarak geliştirilmeye devam edilmektedir. [2]

Zemberek doğal dil işleme kütüphanesine ek olarak, derlem içindeki kelimelerin morfolojik analiz işlemi için açık kaynak kodlu TRmorph kütüphanesinden ve dil tespiti için Google tarafından geliştirilen açık kaynak kodlu Compact Language Detector kütüphanesinden yararlanılmıştır. TRmorph kütüphanesi C dili ile yazılmış, 2007 yılından beri geliştirilmeye devam edilen bir kütüphanedir. [3] Dil tespiti için tercih edilen Compact Language Detector kütüphanesi ise 83 farklı dili olasılık tabanlı yöntemlerle tespit edebilmektedir. [4]

2 Derlemin Hazırlanması

Bu çalışmada anlatılan Sağlık Bilimleri Türkçe Derlemi, tamamı internet üzerinden erişilebilir olan akademik dergi web sitelerinden alınan makalelerden oluşmaktadır. Derlem, sağlık bilimleri alanında yayımlanan 94 farklı derginin

arşivlerinden oluşturulmuştur. Dergilerin arşivleri hazırlanan bir betik ile çözümlenmiş ve dergilerin sitesindeki makalelerin tam metin dosyaları indirilmiştir. İndirilen tam metin dosyalarının içindeki metinler çıkartılmış, çıkarılan metinlerin dili tespit edildikten sonra yalnızca Türkçe olan makaleler kullanılmıştır. Çıkartılan metinler açık kaynak kodlu bir yazılım olan Zemberek kütüphanesinin cümle analiz motoru kullanılarak cümlelere ayrılmıştır. Elde edilen cümleler sözcük türü işaretleme (İng. *part-of-speech tagging*) işlemiyle işaretlenmiş ve kelimelerin kelime kökleri bulunmuştur. Sözcük türü işaretleme işlemine ek olarak kelimelerin morfolojik yapıları TRmorph kütüphanesiyle tespit edilmiş ve derleme eklenmiştir.



Şekil. 1. Derlemin hazırlanma aşamaları.

Derlemin hazırlanması aşamasında yapılan tüm işlemler Şekil 1'de özetlenmiştir.

2.1 Makale Dosyalarının İndirilmesi

Sağlık Bilimleri Türkçe Derlemi, sağlık bilimleri alanında DergiPark'ta açık erişimli olarak Türkçe yayımlanan dergiler arasından seçilen 94 derginin internet sitelerinde bulunan arşivlerinden alınan 12.930 makale kullanılarak oluşturulmuştur. Derlem hazırlanırken kullanılan dergilerin listesi Ek A'da verilmiştir. Dergilerin arşiv sayfaları dergilere özel hazırlanan betikler kullanılarak çözümlenmiş, dergilerin arşivlerinde bulunan tüm makalelerin tam metin dosyaları indirilmiştir. İndirilen tam metin dosyalarının dosya formatları kontrol edilmiş,

sadece PDF formatında olan tam metin dosyaları kullanılarak derlem hazırlanmıştır.

2.2 Makale Dosyalarının İşlenmesi

Sağlık bilimleri alanında yayımlanan dergilerin arşivlerinde bulunan makalelerinin PDF formatındaki tam metin dosyaları indirildikten sonra bu dosyalar öncelikle Poppler araçları kullanılarak, işlenebilir XML formatına dönüştürülmüştür. Oluşturulan XML metin dosyalarının içindeki bilgiler hazırlanan betikler ile çıkartılmıştır. XML formatındaki tam metin dosyasının orta sayfalarından alınan örnekler kullanılarak makalenin dil tespiti yapılmış, Türkçe olmayan dosyalar ayıklanmıştır.

Türkçe olduğu tespit edilen tam metin dosyalarından oluşturulan XML dosyaları, PHP diliyle hazırlanmış betikler yardımıyla işlenmiş ve makale sayfalarındaki sayfa üst yazıları ile sayfa alt yazıları bu yazıların diğer sayfalarda da tekrar etmesi göz önüne alınarak silinmiştir. Bu işleme ek olarak, makale içinde bulunan ancak kurallı cümle yapısında olmadığı için doğal dil işleme aşamalarında hatalı sonuçlar oluşturacak ve makale dili Türkçe olmasına rağmen farklı dillerde ögeler barındırabilecek bölümler (örn. kaynakça bölümü ve tüm dillerdeki özet ve anahtar kelime bölümleri) de silinmiştir. İşlemler sonucunda makalelerin içindeki tüm ek yazıların silinmesi sadece makale metninin derleme eklenmesi amaçlanmıştır.

2.3 Derlemin Oluşturulması

Önceki bölümde anlatılan şekilde ayıklanan metinler, açık kaynak kodlu bir yazılım olan Zemberek doğal dil işleme kütüphanesi kullanılarak cümlelerine ayrılmıştır (İng. *sentence boundary detection*). Elde edilen cümleler içindeki kelimelere Zemberek-NLP kütüphanesi kullanılarak sözcük türü işaretleme işlemi ve kelime kökü belirleme işlemleri yapılmıştır. Zemberek kütüphanesinden alınan çıktıların doğruluğunun yükseltilmesi amacıyla, kütüphaneye dahil olan sözlüklere ek olarak İngilizce-Türkçe tıp terimleri sözlüğü çalışma sırasında Zemberek'e eklenmiştir. Sözcüklerin morfolojik analizi için açık kaynak kodlu olan TRmorph kütüphanesi kullanılmıştır. Alınan sonuçlar birleştirilmiş ve Sağlık Bilimleri Türkçe Derlemi oluşturulmuştur.

Derlem içinde cümleleri oluşturan kelimelerin ham hali, kelimelerin kökü, cümle içindeki görevi ve morfolojik detayları yer almaktadır. Kelime kökleri, kelimelerin cümle içindeki görevi ve kelimelerin morfolojik detayları seçilirken cümle içinde belirsizlik analizi yapılmış, seçilen değerler belirsizlikler çözümlendikten sonra derleme alınmıştır. Derlem içinden örnek bir kısım Tablo 1'de gösterilmiştir.

Sağlık Bilimleri Türkçe Derlemi'nde bu kapsamda 23.271.623 öge bulunmaktadır. Çalışma kapsamında oluşturulan Türkçe derlem akademik çalışmalarda kullanılmak üzere bu bildirinin kaynak gösterilmesi kaydıyla erişime açıktır. Derlemi kullanmak için yapılması gereken işlemler sonraki bölümde açıklanmıştır.

Kelime	Kök	Görev	Morfoloji
pansitopeni	pansitopeni	Noun	
genellikle	genellikle	Adverb	genel<Adj><0><N><lik><N><ins>
yoğun	yoğun	Adjective	yoğun<Adj>
kemik	kemik	Noun	kemik<N>
iliği	ilik	Noun	i<Num:rom><0><N><lik><N><p3s>
fibrozisi	fibrozis	Noun	
ile	ile	Conjunction	i<Num:rom><0><N><ins>
ilişkilidir	ilişki	Verb	ilişki<N> <Adv><0><N> <0><V><cpl:pres><3s><dir>

Tablo. 1. Sağlık Bilimleri Derlemi içinden bir bölüm.

3 Derlem Kapsamı Testi

Derlem bazlı yapılan dilbilim çalışmalarında kullanılan derlemler aracılığıyla dilin kullanımı ile ilgili olarak çıkarımlar yapılabilse de, yapılacak çıkarımların doğru olması için kullanılan derlemlerin dil dağarcığını kapsama oranı yüksek olmalıdır. [5] Bu nedenle, oluşturulan derlemin kapsamının yüksek olması amaçlanmıştır.

Bu çalışmada oluşturulan derlemin kapsamının tespiti için Türkiye sağlık bilimleri ulusal veri tabanından rastgele alınmış 198 makale ile kapsam testi yapılmıştır. Ulusal sağlık bilimleri veri tabanından alınan makalelere derlemi oluştururken yapılan ön işlemler (İng. *preprocess*) uygulanmış, elde edilen kelimeler hem oluşturulan derlem içinde hem de Türkçe Vikisözlük içinde aranmıştır. Kapsam testi sonucunda, bu çalışmada anlatılan sağlık bilimleri derleminin ulusal sağlık bilimleri veri tabanından alınan 198 makale içindeki 310.871 kelimenin %88'ini kapsadığı tespit edilmiştir. İnternetten erişilebilen, 328.409 kelime içeren güncel bir Türkçe sözlük olan Vikisözlük'ün ise aynı test verisinin ancak %58'ini kapsadığı görülmüştür.

Gerçekleştirilen karşılaştırmalı kapsam testi sonucunda çalışmanın çıktısı olan derlemin sağlık bilimleri alanında, genel maksatlı güncel bir Türkçe sözlüğe kıyasla %52 oranında daha geniş kapsama sahip olduğu gözlenmiştir.

4 Derlemin Kullanılması

Derlemi kullanmak isteyen araştırmacıların isimlerini, unvanlarını, kurumlarını ve yapılacak çalışmanın amacını acikveri@yt.com.tr e-posta adresine bildirmeleri gerekmektedir. Oluşturulan derlemin anasayfası `http://acikveri.yt.com.tr/saglik/derlem` olarak belirlenmiştir.

5 Sonuçlar

Bu çalışma kapsamında, sağlık bilimleri alanında yapılan akademik çalışmalardan oluşan 23.271.623 tekrarlı öge içeren bir derlem oluşturulmuştur. Oluştu-

rulan derlemde yer alan ögeler kelime köklerine veya sözcük türü etiketine göre tekilleştirilmemiş, tüm ögeler alınan kaynakta geçtiği haliyle derleme eklenmiştir. Derlem kelime kökü (İng. *lemma*), sözcük türü etiketleri (İng. *part-of-speech tags*) ve kelimelerin morfolojik analizini içermektedir. Oluşturulan derlem özellikle sağlık bilimleri alanında yapılacak çalışmalar olmak üzere tüm akademik çalışmalarda bu bildiri kaynak gösterilerek kullanılabilir.

Notlar: Bu çalışma Yonca Teknoloji'nin TÜBİTAK 1507 KOBİ ArGe başlangıç destek programı altında desteklenen 7160877 numaralı projesi kapsamında gerçekleştirilmiştir.

Kaynaklar

1. Karaoğlu, S.: Türkçe Çevirimiçi Derlemler Üzerine. KMÜ Sosyal ve Ekonomik Araştırmalar Dergisi. 16, 181–188 (2014)
2. Zemberek-NLP Projesi Github Sayfası, `https://github.com/ahmetaa/zemberek-nlp`
3. Çöltekin Ç.: A Freely Available Morphological Analyzer for Turkish In Proceedings of the 7th International Conference on Language Resources and Evaluation (2010)
4. Compact Language Detector 2 Github Sayfası, `https://github.com/CLD2Owners/cld2`
5. Biber, D.: Representativeness in Corpus Design. Literary and Linguistic Computing. 8(4), 243–257 (1993)

Ek A Derlemde Kullanılan Dergiler

Dergi Adı	Makale Sayısı
Atatürk Üniversitesi Diş Hekimliği Fakültesi Dergisi	709
Acta Odontologica Turcica	696
Dicle Tıp Dergisi	683
Ege Tıp Dergisi	539
Turgut Özal Tıp Merkezi Dergisi	520
Türk Pediatri Arşivi	507
Journal Of Anatolia Nursing And Health Sciences	490
Fırat Tıp Dergisi	444
Süleyman Demirel Üniversitesi Tıp Fakültesi Dergisi	372
Kocatepe Tıp Dergisi	361
Journal Of Experimental And Clinical Medicine	307
Çukurova Üniversitesi Tıp Fakültesi Dergisi	305
Cumhuriyet Medical Journal	300
Akademik Gastroenteroloji Dergisi	296
Cerrahpaşa Tıp Dergisi	277
Ondokuz Mayıs Üniversitesi Diş Hekimliği Fakültesi Dergisi	268

Ankara Üniversitesi Tıp Fakültesi Mecmuası	268
Florence Nightingale Hemşirelik Dergisi	258
Journal Of Istanbul University Faculty Of Dentistry	254
Journal Of Contemporary Medicine	224
Konuralp Tıp Dergisi	218
Tıp Eğitimi Dünyası	209
Mustafa Kemal Üniversitesi Tıp Dergisi	205
İstanbul Tıp Fakültesi Dergisi	203
Gümüşhane Üniversitesi Sağlık Bilimleri Dergisi	190
Zeynep Kamil Tıp Bülteni	189
Balkan Medical Journal	185
Cumhuriyet Dental Journal	178
Sakarya Tıp Dergisi	173
Süleyman Demirel Üniversitesi Sağlık Bilimleri Dergisi	163
Marmara Medical Journal	144
Hacettepe Üniversitesi Hemşirelik Fakültesi Dergisi	141
Kırıkkale Üniversitesi Tıp Fakültesi Dergisi	132
Dokuz Eylül Üniversitesi Tıp Fakültesi Dergisi	131
Ankara Medical Journal	124
Acta Oncologica Turcica	119
Yoğun Bakım Hemşireliği Dergisi	119
Osmangazi Tıp Dergisi	117
Türk Onkoloji Dergisi	109
Gaziantep Medical Journal	108
Göğüs-Kalp-Damar Anestezi Ve Yoğun Bakım Derneği Dergisi	105
Clinical And Experimental Health Sciences	102
International Journal Of Basic And Clinical Medicine	93
Marmara Pharmaceutical Journal	92
Türk Fizyoterapi Ve Rehabilitasyon Dergisi	89
Sağlık Bilimleri Ve Meslekleri Dergisi	84
Koşuyolu Kalp Dergisi	73
Bozok Tıp Dergisi	70
İstanbul Bilim Üniversitesi Florence Nightingale Tıp Dergisi	69
Medicine Science	67
Selcuk Dental Journal	59
Ordu Üniversitesi Tıp Dergisi	55
Celal Bayar Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi	53
Atatürk Üniversitesi Tıp Dergisi	48
Erciyes Üniversitesi Sağlık Bilimleri Fakültesi Dergisi	44
Uluslararası Klinik Araştırmalar Dergisi	44
Mersin Üniversitesi Sağlık Bilimleri Dergisi	44
Cumhuriyet Hemşirelik Dergisi	44

Mehmet Akif Ersoy Üniversitesi Sağlık Bilimleri Enstitüsü Dergisi	41
Turkish Journal Of Clinics And Laboratory	38
Anadolu Kliniği Tıp Bilimleri Dergisi	32
Medical Sciences	31
Deneysel Tıp Araştırma Enstitüsü Dergisi	30
Düzce Üniversitesi Tıp Fakültesi Dergisi	25
İstanbul Bilim Üniversitesi Florence Nightingale Transplantasyon Dergisi	22
Journal Of Medical Updates	21
İzmir Katip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi	20
Online Türk Sağlık Bilimleri Dergisi	20
Adıyaman Üniversitesi Sağlık Bilimleri Dergisi	18
Ortadoğu Tıp Dergisi	18
Turkish Journal Of Family Medicine And Primary Care	17
Nefroloji Hemşireliği Dergisi	15
Aile Hekimliği Ve Palyatif Bakım	13
Samsun Sağlık Bilimleri Dergisi	13
Sürekli Tıp Eğitimi Dergisi	9
Pediatric Practice And Research	9
Prusias Tıp Dergisi	8
Acta Medica Alanya	7
Archives Of Clinical And Experimental Medicine	6
Kahramanmaraş Sütçü İmam Üniversitesi Tıp Fakültesi Dergisi	6
European Journal Of Health Sciences	6
Balıkesir Medical Journal	5
Namık Kemal Tıp Dergisi	5
Güncel Dermatoloji Dergisi	4
Gazi Sağlık Bilimleri Dergisi	4
Tıp Araştırmaları Arşivi	3
Turkish Journal Of Medical Sciences	3
Medical Genetics	3
Journal Of Anatolian Medical Research	2
Ankara Eğitim Ve Araştırma Hastanesi Tıp Dergisi	2
Ege Üniversitesi Hemşirelik Fakültesi Dergisi	1
Erciyes Üniversitesi Sağlık Bilimleri Dergisi	1
İbni Sina Tıp Bilimleri Dergisi	1
Gazi Üniversitesi Diş Hekimliği Fakültesi Dergisi	1
Toplam	12930

Tablo. A.1. Sağlık Bilimleri Derlemi oluşturulurken kullanılan dergilerin listesi.