TÜRKÇEDE SIKÇA "YANLIŞ YAZILAN" SÖZCÜKLERİN BÜYÜK ÖLÇEKLİ DERLEMLERDE TESPİT EDİLMESİ

Ahmet ARSLAN[a], İlknur CİVAN[b], Ümit Deniz TURAN[b]
[a] Bilgisayar Mühendisliği Bölümü, Eskişehir Teknik Üniversitesi, Eskişehir
[b] Yabancı Diller Eğitimi Bölümü, Anadolu Üniversitesi, Eskişehir

ÖZET

Türk Dil Kurumu 2010 yılında *Sıkça Yapılan Yanlışlara Doğrular* sözlüğü yayınlamıştır. Bu sözlükte yaklaşık 500 adet yanlış-doğru çifti bulunmaktadır. Türk Dil Kurumu, çoğu yabancı kökenli olan bu sözcüklerin ölçünlü hale gelmesini sağlamak amacıyla bu sözlüğü sunmuştur. Bu çalışmanın amacı Bilgi Erişimi ve Metin Sınıflandırması için yazılı basın metinlerinden oluşturulmuş Türkçe derlemlerde ve ders kitaplarında anılan sözlükte belirtilen yazım yanlışlıklarının ne sıklıkta yapıldığını tespit etmektir. Tüm derlem üzerinde gerçekleştirilen bu işlem, alt kırılımlarda (spor, sağlık, siyaset, ekonomi, vb.) tekrarlanarak alan/kategori özelinde de istatistikler sunulmaktadır. Tüm derlem seviyesinde en çok yapılan yazım yanlışlıkları: *itibariyle* (*itibarıyla*), *döküman* (*doküman*), *orjinal* (*orijinal*) gibi sözcükler olurken; spor kategorisinde ise *cimnastik* (*jimnastik*), *antreman* (*antrenman*) gibi alana özel sözcükler tespit edilmiştir. Sağlık alanında yazılmış ders kitaplarında *tetanoz* (*tetanos*), *menapoz* (*menopoz*), *ötenazi* (*ötanazi*), ve *sezeryan* (*sezaryen*) sözcükleri saptanmıştır. Sosyoloji ders kitaplarında ise *proleterya* (*proletarya*) ve *psikiyatrist* (*psikiyatr*) sözcükleri dikkat çekmektedir. Türk Dil Kurumu, parantez içindeki sözcükleri doğru olarak önermektedir.

Türk Dil Kurumu'nun yazım yanlışlıkları olarak belirlediği bu sözcüklerin gazete, kitap, makale vb. metinlerde gerçekten de kullanıldığını göstererek farkındalık yaratmak, doğru olarak önerilen yazımların Türkçe öğretiminde öne çıkarılması ya da otomatik imla düzeltme sistemlerinin kullanılması önerilmektedir. Bunun yanı sıra bu sözcükler, yabancı dillerden ödünçlenerek alındığından bu konuda dilbilimsel bir artalan ve Türk Dil Kurumu'nun yaklaşımına ilişkin kısa bir değerlendirme de sunulacaktır.

IDENTIFICATION OF THE MOST FREQUENTLY "MISSPELLED TURKISH WORDS" IN LARGE-SCALE CORPORA

ABSTRACT

The Turkish Language Association released a dictionary for frequently misspelled words in 2010. This dictionary contains approximately 500 wrong-correct words pairs. In this study, the aim is to present the observed frequencies of these misspelled Turkish words in Information Retrieval and Text Classification datasets as well as textbooks written on various domains. The results show that these words are in fact attested in news articles and textbooks. Some misspelled words in the whole corpus are as follows: *itibariyle* (*itibarıyla*), *döküman* (*doküman*), *orjinal* (*original*). The words in parentheses are suggested as correct versions. When subcategories are analyzed, *cimnastik* (*jimnastik*), *antreman* (*antrenman*) are observed in the sports news; *tetanoz* (*tetanos*), *menapoz* (*menopoz*), *ötenazi* (*ötanazi*), and *sezeryan* (*sezaryen*) in health textbooks; *proleterya* (*proletarya*) and *psikiyatrist* (*psikiyatr*) in sociology textbooks.

In the current study, the most frequently observed misspelled words are examined also from a linguistic perspective. Since these so-called misspelled words are all borrowed from various languages, we will provide a brief background on borrowing and a brief evaluation on the suggestions made by the Turkish Language Association.

GİRİŞ

Dil canlı bir varlıktır ve zaman içerisinde farklı nedenlerle değişime uğrayabilmektedir. Bu değişimin içerisinde başka dillerden sözcük ödünçleme ve o dile yerleştirmenin de bir payı vardır çünkü diller toplumlar gibi birbirlerini etkilemektedir. Dil etkileşiminin bir sonucu da başka bir dilden alınan sözcüklerdir. Bu sözcük oluşturma sürecine ödünçleme denir. Ödünçleme sözcükler, tarihsel açıdan ortak bir kök dilden gelen kökteş sözcükler dışında başka bir dilden alınan sözcüklerdir. Her dilde bulunan çekirdek/temel sözcük dağarcığı, örneğin sayılar, akrabalık terimleri, gibi yerli dilde bulunan sözcüklerdir; bunlar çoğunlukla ödünç alınmazlar. Öte yandan, temel sözcük varlığının dışında araştırmacılar, her dilde sözcük ödünçlemenin çok yaygın olduğu görüşündedirler (Örn. Bkz. Treffers-Daller, 2010 ve içindeki kaynakça). Johanson (1998) ödünçleme için "kopyalama" terimini kullanarak Farsça ve Türkçenin etkileşimini incelemektedir. Osmanlı döneminde eğitimli kesim prestijli dil sayılan Arapça ve Farsçadan etkilenerek onların aracılığıyla çok fazla sözcük almış ve Osmanlı Türkçesi, Türkçe, Arapça ve Farsça ve diğer dillerden ödünçlenen sözcükleri içeren bir dile dönüşmüştür. Türkçe, tarihte denizcilikte ileri olan Venedikliler aracılığıyla İtalyanca ve Yunancadan denizcilik terimleri ödünçlemiştir (Winter, 1960). İstilalar da dil etkileşimi ve ödünçleme konusunda en önemli sosyo-politik unsurlar arasındadır (Sankoff, 2002). Yüzyıllar boyu İber yarımadasını işgal etmiş olan Magripliler, İspanyolcada bugünde kullanılan pek çok Arapça sözcük bırakmışlardır. Görüldüğü gibi, ödünçleme, dillerin etkileşimi sonucu ortaya çıkar. Göçler, coğrafi yakınlık, toplumlararası ticaret, dinsel etmenler, bir dilin eğitim dili olarak prestijli sayılması, bilim ve teknoloji alanında yeni buluşlar ve keşifler, küreselleşme, kültürel etmenler gibi dillerin etkileşimini sağlayan diğer unsurlar sonucunda her dil başka bir dilden sözcükler ödünç alabilir. Örneğin, "*avakado*" yeni bir sebze türü olarak İspanyolların Güney ve Orta Amerika'yı fethetmesi sonucu yerel Nahuatl dilinden İspanyolcaya ve sonra İngilizceye geçmiştir. Daha sonra Türkçede de kullanılmaya başlanmıştır. Benzer şekilde kültürel etkileşim sonucu "*spagetti*" bir İtalyan yemeği olarak İngilizce ve Türkçede de kullanılmaktadır. Ödünçleme, her ne kadar dilde eleştirilen bir durum olsa da doğal bir süreçtir. Bu ödünçleme süreci çoğu zaman çeviri ya da doğrudan dile girmesi işlemiyle gerçekleşebilir (Weinreich, 1953; Balcı, 2017). Bu aynı zamanda dilleri geliştiren de bir süreçtir. Burada önemli olan bir diğer nokta da ana dili konuşucuların sözcük alınan yabancı dili tanıma ölçüleridir çünkü alıntı sözcüklerin hedef dile geçerken uygun hale getirilmesinde verici dilin özelliklerini bilme de rol oynar (Bloomfield, 1933; Weinreich, 1953, vd.). Dillerin bunun gibi özelliklerini incelemek üzere farklı çalışmalar yıllardır yapılmakta olup (Zimmer, 1985; Friesner, 2009) son yıllardaki teknolojik gelişmelerle büyük ölçekli veriler üzerinde çalışabilmek ve bu verileri işlemek konusunda da gelişmeler gözlenmektedir (Haspelmath ve Tadmor, 2009).

Bu konuda yapılan akademik çalışmaların yanında bazı kurumlar da önemli rol oynamaktadır. Dilde birliğin sağlanması amacıyla ülkemizde çalışmakta olan kurum Türk Dil Kurumu'dur. Çalışmalarını Türk Dili üzerinde araştırmalar ve Türk Dilinin güncel sorunları olmak üzere iki ana alanda yürütmektedir. Her ne kadar bu amaca yönelik pek çok çalışma veya sözlük oluşturulsa da yine de Türkçenin sözcük varlığındaki özellikle yabancı dilden ödünç alınan sözcükler farklı ana dili konuşucuları tarafından farklı sesletilmekte ve farklı yazılabilmektedir.

Bu yabancı kökenli sözcüklerin ve özelikle daha eskiden alınmış olanların Türkçedeki söylenişleri ve yazılışları Türkçenin kurallarına uyumlu hale dönüşmüştür. Aslında yabancı sözcüklerin ölçünlü biçimleriyle Türkçede kullanılması söyleyişte ve yazımda birliğin sağlanması açısından önemlidir ve Türkçe karşılıkları varsa, bu karşılıkların kullanılması doğru bir çözüm olabilir.

Ancak Türkçede karşılığı yaygınlaşmamış sözcüklerin kullanılması durumunda bu sözcüklerin farklılık gösterebildiği görülmektedir. Bu tür değişkenliklerin önüne geçerek bir standart belirlemek amacıyla oluşturulan *Sıkça Yapılan Yanlışlara Doğrular Sözlüğü*, Türk Dil Kurumunun yetmiş sekizinci kuruluş yıl dönümü törenleri kapsamında 12 Temmuz 2010 günü kamuoyuna açılmıştır. Bu sözlükte yaklaşık 500 adet yanlış-doğru olarak önerilen sözcük çifti bulunmaktadır (http://sozluk.gov.tr). Çalışmada da görüldüğü üzere bu sözlükte yanlış olduğu belirtilen sözcüklerin, gerçekte anadil konuşucuları tarafından yazılı Türkçe metinlerde kullanıldığı gözlenmektedir. Bu çalışmanın amacı, hangi sözcüklerde daha sık değişkenlik olduğunu ve yazım yanlışı yapıldığını tespit etmek ve bunun sebeplerinin neler olabileceğini belirlemektir. Bu amaçla yapılan bu araştırmada bu yanlışlar otomatik tespit edilerek; frekanslarına göre büyükten küçüğe doğru sıralanarak incelenmektedir. Çalışmamızda "yazım yanlışı" ifadesinden herhangi bir yazım yanlışı değil, TDK'nın *Sıkça Yapılan Yanlışlara Doğrular Sözlüğünde* yanlış olduğu belirtilen sözcükler anlaşılmalıdır.

Araştırmanın Amacı ve Önemi

Bu çalışmanın ana amacı TDK'nın "*Sıkça Yapılan Yanlışlara Doğrular*" sözlüğünde verilen sözcüklerin Türkçe derlemlerde ne sıklıkla gözlemlendiğinin tespit edilmesidir. En sık gözlemlenen yanlışlar hem dilbilimsel hem de hesaplamalı metin işleme perspektifinden incelenmiştir. Bu çalışmanın katkıları aşağıdaki şekilde özetlenebilir.

1. TDK'nın "Sıkça yapılan yanlışlara doğrular" sözlüğündeki yazım yanlışlarının haber metinleri, açık erişimli ders kitapları gibi içeriklerde gerçekten de büyük oranda yapıldığını göstererek farkındalık yaratmak.
2. Söz konusu sözcüklerin neden yanlış yazıldıklarını dilbilimsel çerçevede açıklamaya çalışmak.
3. Çalışma süresince tespit edilen iki yazım yanlışını (*biodizel-biyodizel* ve *hemşeri-hemşehri*) TDK yetkililerine bildirmek.

Çalışmanın geri kalan bölümleri şu şekilde düzenlenmiştir. 2. bölümde, incelenen derlemler, kullanılan açık kaynak kodlu yazılım ve frekans çeşitleri açıklanmaktadır. Deneysel bulgular ve yapılan çözümlemeler 3. bölümde sunulmuştur. 4. bölümde ise sonuç ve gelecekteki çalışmalara yönelik olası tavsiyeler verilmektedir.

YÖNTEM

Bu bölümde deneysel metodoloji anlatılmaktadır.

İncelenen Derlemler

Çalışmamızda ikisi metin sınıflandırması, biri bilgi erişimi için oluşturulmuş olmak üzere üç Türkçe derlem kullanılmıştır. Bir başka deyişle, umuma açık mevcut derlemler kullanılmış olup yeni bir derlem oluşturulmamıştır. Bahsi geçen derlemlerin hepsi yazılı basın çevrimiçi gazete Web sitelerinden toplanmıştır. Veri kümeleri haber metinlerinden oluştuğu için, içeriği yayınlayan gazetenin sınıflandırması kullanılmıştır. Örneğin, spor haberleri spor kategorisi, ekonomi haberleri ekonomi kategorisi altında toplanmıştır. Aslına bakarsak, çevrimiçi haber siteleri kategorisi belli metin elde etmek için elverişli ve müsaittir. Çalışmada kullanılan derlemler aşağıda tanıtılmaktadır:

- **Milliyet Koleksiyon** www.milliyet.com.tr çevrimiçi gazetesinden 2001-2005 yılları arasında toplanmış 408.305 adet haber metni ve köşe yazılarından Bilkent Üniversitesi bünyesinde oluşturulmuştur (Can vd., 2008).
- **42 Bin Haber Veri Kümesi** 13 haber grubuna (kategori) ait, toplamda 41.992 adet haberin metninden Yıldız Teknik Üniversitesi bünyesinde oluşturulmuştur (Yıldırım ve Atık, 2013).
- **TTC-3600** Hürriyet, Posta, İhlas Haber Ajansı, HaberTürk, ve Radikal çevrimiçi gazetelerinden Mayıs-Temmuz 2015 tarihleri arasında toplanmış 3.600 adet haber metninden Celal Bayar Üniversitesi bünyesinde oluşturulmuştur. Bu derlemin adı olan TTC, Türkçe Metin Sınıflandırmanın İngilizcesi olan "*Turkish Text Categorization*" ibaresinin baş harflerinden oluşan kısaltmadır (Kılınç vd., 2017).

Derlemlere ait kategoriler ve belge sayıları Tablo 1'de verilmiştir. Derlemler farklı dönemlerde farklı kaynaklardan toplandıkları için kategoriler çeşitlilik göstermektedir. Haber siteleri arasında ortak kullanılan kategoriler Ekonomi, Kültür/Sanat, Politika/Siyaset, ve Spor'dur. En çok belgeye sahip derlem 400 binin üzerinde belge ile Milliyet Koleksiyon iken, en çok kategoriye sahip olan 42 Bin Haber veri kümesidir. TTC-3600 ise belgelerin kategorilere eşit dağıldığı tek derlem olup, diğer iki derlemde dengesiz dağılım söz konusudur.

Tablo 1. Derlemlerin belge sayıları.

	TTC-3600	42 Bin Haber	Milliyet Koleksiyon
Astroloji			1824
Dünya		3724	68575
Ekonomi	600	3265	73776
Genel		6673	
Güncel		5847	20983
Kültür/Sanat	600	1155	159
Magazin		2792	10003
Planet		1953	
Politika/Siyaset	600	1849	30164
Sağlık	600	1383	
Spor	600	9997	81766
Teknoloji	600	771	
Türkiye		1939	62679
Yaşam		644	32316
Yazar(lar)			25684
Toplam	3600	41992	405570

Apache Lucene

TDK tarafından önerilen sözlüğe göre yazım yanlışlarının gözlemlenme frekanslarını elde etmek için açık kaynak kodlu yazılım olan Apache Lucene (Białecki vd., 2012) kullanılmıştır. Apache Lucene[1] her ne kadar bilgi erişimi kütüphanesi olsa da büyük veri kümelerinden frekansların çıkarılması için de kullanılabilmektedir. Örneğin Arslan vd. (2018) yarım milyar Web sayfası içinde en çok geçen emojileri Apache Lucene kullanarak elde etmişlerdir.

[1] http://lucene.apache.org

Frekans Çeşitleri

Bir sözcüğe ait frekansın iki türlü hesaplaması yapılmıştır. Birisi "*olduğu gibi geçtiği*" frekans, diğeri ise "*ile başlayan*" terim frekansıdır. *Hemşeri* yazım yanlışı (doğrusu *hemşehri*) dikkate alındığında, *hemşerilerim*, *hemşerim* gibi değişkenler "*olduğu gibi geçtiği*" tipi frekans hesabında dikkate alınmaz iken, "*ile başlayan*" frekans tipinde dikkate alınmaktadır. Türkçenin sondan eklemeli yapısı dikkate alındığında sıralama kriteri olarak "*ile başlayan*" tipi frekansı kullanmanın uygun olduğu kanaatine varılmıştır. Ancak birbirinin karakter dizisi olarak alt kümesi olan kelimelerin bu frekans hesabında yanlış eşleşmelere neden olduğu unutulmamalıdır. Örneğin *Yunanlı-Yunan*, *artis-artist*, *eşyalar-eşya* gibi.

Sözlükte yapılan sadeleştirmeler

TDK sözlüğünde bir takım sadeleştirmeler yapılmıştır. Sözlükte okunuş farklılıklarını içeren girdiler vardır. Bunların yazımı farklılık göstermediği için elenmiştir. Bazı girdiler ise bağlama bağımlıdır. Örneğin *Eğridir-Eğirdir* çifti sözlükte yer almasına rağmen bizim çalışmada elenmiştir. Çünkü "*eğridir*" sözcüğü şehir ismi olarak yanlıştır ama başka bir bağlamda doğru olabilir. Çalışmada sadece bağlamdan bağımsız yanlışlar kullanılmıştır.

DENEYSEL BULGULAR VE ÇÖZÜMLEME

Bu bölümde TDK'nın sözlüğünde verilen yanlış-doğru çiftleri, Milliyet Koleksiyon veri kümesindeki gözlemlenme sıklıklarına göre büyükten küçüğe sıralanarak verilmiştir. Çalışmada üç veri kümesi incelenmesine rağmen, hem benzer sonuçlar elde edilmesinden dolayı hem de üç veri seti için ayrı ayrı verilecek sonuçların karmaşa yaratması sebebiyle sayıca en çok belgeyi (400 binden fazla) içeren Milliyet Koleksiyon baz alınmıştır.

Bir derlem üzerinden bir sözcüğe ait, *belge frekansı* ve *sözcük frekansı* olmak üzere iki farklı frekans hesaplamak mümkündür. Metin sınıflandırma ve bilgi erişimi (Manning vd., 2008) terim ağırlıklandırma yöntemleri de bu iki istatistiği kullanmaktadır (Hiemstra, 2000). Bunlardan ilki o sözcüğün kaç adet belgede gözlemlendiği belge frekansıdır. Bu frekans çeşidinin birimi belge sayıdır. Diğer frekans çeşidi ise sözcük frekansıdır ve tüm derlem içinde o sözcüğün toplamda kaç kere geçtiğini gösterir. Bu frekans çeşidinin birimi ise sözcük sayısıdır. Her iki frekans tipi verilen bir sözcük ve derlem çifti için hesaplanabilir.

Milliyet Koleksiyon üzerinden elde edilmiş en sık yapılan 45 yazım yanlışı ve bu sözcüklerin doğruları Tablo 2.'de verilmiştir. Tablodaki satırlar yanlış sözcüklerin "ile başlayan" belge frekansına (tabloda * ile gösterilmiştir) göre büyükten küçüğe doğru sıralanmıştır. Bir sözcüğün "ile başlayan" sözcük frekansını hesaplamak teknik olarak zor olduğu için mevcut çalışmada kullanılmamıştır.

Tablo 2.'den anlaşılacağı üzere, en sık gözlemlenen yanlışlar çoğunlukla ödünç sözcüklerdir. Bu yapılan yazım yanlışının daha çok kaynak dildeki okunuşlarından kaynaklanabileceği düşünülmektedir. Örn: *antrenman* (Fransızca), *ataş* (Fransızca), *dakika* (Arapça), *şoför* (Fransızca), *ızdırap* (Arapça), *fayton* (Fransızca). Bu durumda, TDK kaynak dildeki okunuşlarını esas almaktadır. Yani eğer, ana dili konuşucuları kaynak dili bilmiyorlarsa bu sözcükleri Türkçenin ses kurallarına uyumlu hale getirebilmekte ve Türkçe sesletildiği biçimde yazılan bir dil olduğu için bunu yazım biçimine de yansıtabilmektedirler.

Tablo 2. Milliyet derleminde en sık gözlemlenen 45 yazım yanlışı ve doğrusu.

Yanlış	**Sözcük Frekans**	**Belge Frekans**	**Belge* Frekans**	**Doğru**	**Sözcük Frekans**	**Belge Frekans**	**Belge* Frekans**
itibariyle	19450	15571	15577	itibarıyla	4038	2826	2826
hıristiyan	3965	2750	3471	hristiyan	304	212	287
eşyalar	549	469	2527	eşya	3437	2392	6092
mönü	1710	950	1731	menü	67	65	216
makina	738	524	1583	makine	1936	1445	4820
yunanlı	1327	963	1555	yunan	9666	4900	13674
hemşeri	102	81	674	hemşehri	70	61	399
mevlüt	783	571	584	mevlut	19	16	24
ünvan	47	43	505	unvan	384	306	3376
tesbit	359	329	463	tespit	18441	14387	17599
hintli	487	366	461	hint	949	698	1234
artis	10	10	451	artist	149	121	434
müdahele	275	253	433	müdahale	12006	8910	15527
acenta	36	30	414	acente	140	101	546
antreman	132	104	398	antrenman	4434	3499	9365
mahçup	318	286	334	mahcup	466	403	414
penbe	298	242	302	pembe	1313	972	1086
sarmısak	327	207	302	sarımsak	232	158	253
ataç	403	228	270	ataş	167	91	564
meslekdaş	14	14	258	meslektaş	61	56	2872
meyva	155	127	253	meyve	3338	2049	3204
yanlız	173	159	245	yalnız	8379	7105	13812
süpriz	177	165	242	sürpriz	8474	6119	7846
sandöviç	195	142	240	sandviç	277	204	324
çoşku	19	19	235	coşku	526	462	5293
cimnastik	267	191	227	jimnastik	348	265	283
psikiyatrist	184	148	218	psikiyatr	261	207	996
laboratuar	108	84	217	laboratuvar	758	648	1828
usül	73	64	201	usul	1876	1374	6167
ıstırap	143	125	182	ızdırap	54	50	77
traş	154	116	179	tıraş	451	262	422
payton	220	155	174	fayton	44	27	79
dakka	90	76	162	dakika	22154	15393	29300
siluet	25	17	152	silüet	4	4	43
kollektif	167	146	148	kolektif	580	509	512
diyaloga	162	148	148	diyaloğa	323	295	295
adele	24	19	147	adale	173	151	13360
motorsiklet	72	60	141	motosiklet	1037	592	1226
restorant	35	33	138	restoran	2394	1568	3539
döküman	16	16	133	doküman	229	214	755
yayınlamak	99	95	128	yayımlamak	90	84	115
fantazi	91	81	125	fantezi	283	220	414
orjinal	103	90	123	orijinal	896	705	898
entellektüel	116	97	121	entelektüel	757	603	747
parlemento	31	31	117	parlamento	3434	2628	8942

Örneğin, Türkçe genellikle sözcük başında gelen ünsüz çiftlerini kabul etmez (Ercilasun, 2013). Dolayısıyla sözcük başında bulunan ünsüz çiftini bölmek amacıyla iki ünsüzün arasında kullanılabilecek ünlü ses türemesi beklenirken bazı sözcüklerde TDK, sözcüğün alındığı kaynak dile uygun yazımı önermektedir. Örn: Yunancadan gelen ve dilimizin ses kurallarına uygun olarak ünsüz çiftini bölen bir ünlü kullanılarak *Hıristiyan* şeklinde sesletilebilen ve yazılabilen sözcük, TDK'ya göre yanlıştır. TDK bu sözcüğün kaynak dil olan Yunanca biçimiyle kullanımını önermektedir. Bunun yanı sıra vatandaşlık ya da ulusal kimlik adlarında kullanılan *-lı* eki de TDK'ya göre sıkça yapılan yanlışlar arasındadır: Örn: *Yunanlı* (Farsça), *Afganlı* (Arapça), *Hintli* (Farsça) gibi.

Arapçadan Türkçeye girmiş olan *tekrar* sözcüğü, kaynak dilde isimdir ve kimi Türkçe konuşucuları bu sözcüğü zarf olarak *tekrardan* biçiminde kullanmaktadırlar. Oysa TDK'ya göre bu sözcük Türkçede hem isim hem zarf olarak kullanılmalıdır. Benzer şekilde Arapçadan gelen *şey* ve *eşya* sözcükleri bağlantılı olup, ikincisi, birincisinin çoğuludur. Kimi ana dili konuşucuları kaynak dilde zaten çoğul olan bir sözcüğe Türkçede çoğul *-lar* ekini getirerek *eşyalar* biçiminde kullanmaktadırlar.

Kategori Özelinde Deneysel Bulgular

Bir önceki bölümde tüm koleksiyon üzerinden hesaplanan sonuçlar verilmiştir. Kategori özelinde sonuçlar incelendiğinde o alana özel kavram ve sözcüklerin ilk sıralarda çıkması ilgi çekicidir. Örneğin, *proletarya* sözcüğünün Sosyoloji alanında yazılmış kitaplarda gözlenmesi gibi. Tablo 3'te beş kategori ve bu kategorilerin karakteristiklerini yansıtan elle seçilmiş beşer adet yanlış-doğru çifti verilmiştir.

Tablo 3. TDK'nın yanlış olarak önerdiği sözcüklerden kategorilerde en sık gözlenenleri.

Spor	antreman antrenman	cimnastik jimnastik	adele adale	testesteron testosteron	kareografi koreografi
Sağlık	tetanoz tetanos	barsak bağırsak	akapunktur akupunktur	sezeryan sezaryen	allerji alerji
Magazin	mönü menü	sandöviç sandviç	sarmısak sarımsak	salomanje salamanje	restorant restoran
Sosyoloji	proleterya proletarya	kollektif kolektif	entellektüel entelektüel	psikiyatrist psikiyatr	laboratuar laboratuvar
Muhasebe	ıskonto iskonto	harfiyat hafriyat	personeller personel	eşyalar eşya	ünvan unvan

Hesaplamalı Metin İşleme Perspektifi

Bilgisayar ile metin işleme açısından bakıldığında bir sözcüğe ait iki farklı yazılış biçiminden hangisinin doğru hangisinin yanlış olduğu çok da önemli değildir. Ancak, makine tek bir harfin farklı yazılmasını bile iki farklı sözcük olarak algılayacaktır. Farklı yazım varyasyonlarının aynı sayılması birçok bilgisayar ile metin anlama uğraşının (Doğal Dil İşleme, Hesaplamalı Dilbilim, Metin Sınıflandırma, Makine Çevirisi, Bilgi Erişimi, Soru Cevaplama Sistemleri) başarımını artıracaktır. Üstelik, bu çalışmada sunulan yanlış-doğru çifti listesi anılan sistemlere çok kolay bir şekilde entegre edilebilir.

SONUÇ VE GELECEK ÇALIŞMALAR

Yapılan çalışmada görülmektedir ki yanlış yazıldığı belirtilen sözcükler genellikle ödünç/alıntı sözcüklerdir. Ödünçlenirken sözcükler bazı dil olaylarına uğrayabilirler.

Yabancı dillerden ödünç sözcük alan her dilde ana dili konuşucuları (özellikle ilgili yabancı dili bilmiyorlarsa) bu sözcükleri ses, biçimbirim, yazım kuralları açısından kendi dillerinin dilbilimsel özelliklerine uydurabilirler (Poplack, 2017; Sankoff, 2002; Haspelmath, 2009). Örneğin, İngilizceden Fransızcaya geçen *weekend* (hafta sonu) sözcüğü İngilizcedeki kurallar gereği cinsiyet artikeli bulunmamasına karşın Fransızcada her adın mutlaka bir cinsiyeti olması gereğinden hareketle *le weekend* olarak uyarlanmıştır (Haspelmath, 2009: 42). Ödünç sözcüklerin ana dile tam uyum sağlaması uzun bir zaman alan karmaşık bir süreçtir (Haspelmath, 2009). Örneğin Fransızcadan 20. yüzyılda ödünç alınmış olan *garage* (*garaj*) sözcüğü bir değil üç şekilde sesletilebilir: /ˈgaraː(d)ʒ/ /ˈgarıdʒ/ /gəˈraːʒ/. Bunun nedeni de *jimnastik* sözcüğünde bulunan ilk ses olan /ʒ/ sesinin Türkçede olduğu gibi İngilizcede de bulunmamasıdır. Öyleyse, yabancı bir dilden giren sözcüklerin sesletimi bazen ana dili kuralları doğrultusunda değişebilir ya da kaynak dildeki haliyle sesletilebildiği gibi bu iki seçenek aynı anda değişkenlik göstererek kullanılabilir. Doyasıyla, yabancı dilden geçen sözcükler aynı zaman içinde değişkenlik gösterebilir. Bilindiği gibi *jimnastik* sözcüğünün ilk sesi de öz Türkçe sözcüklerde bulunmadığı için ana dili konuşucuları bunu Türkçeleştirerek *cimnastik* olarak ya da özgün şekilde *jimnastik* olarak sesletip yazabilirler.

Görüldüğü gibi, yabancı dilden ödünç alınan sözcükler, ana dilin dilbilimsel ve imla kurallarına uygun hale getirilebilir ya da kaynak dildeki haliyle kullanılması tercih edilebilir. Bir başka deyişle, ödünçlenen sözcüklerde kaynak dile sadık kalındığı da görülür. Bu süreç dil planlamasıyla bağlantılı olabilir. Çalışmamızda TDK'nın önerilerine göre sözcüklerin doğru olarak kabul edilen hallerinde daha çok kaynak dildeki okunuşlarının temel alındığı görülmektedir. Öte yandan TDK, farklı değişkenleri olan yabancı temelli sözcüklerin gerçek hayattaki kullanımlarını izleyerek çoğunluğun kullandığı şekli önermeyi de diğer bir yol olarak izleyebilir.

Çalışma süresince, TDK'nın *Sıkça Yapılan Yanlışlara Doğrular* sözlüğünde yer almayan iki yeni sözcük keşfedilmiştir: (*biodizel-biyodizel* ve *hemşeri-hemşehri*). Bu iki sözcük katki@tdk.gov.tr e-posta adresi aracılığıyla ilgili yetkililere bildirilmiştir. Yakın zamanda bu çalışmanın katkısı olan bu iki yazım yanlışının sözlüğe eklenmesi beklenmektedir.

Gelecek Çalışmalar

Bu yazım yanlışlarının otomatik olarak düzeltilmesinin Türkçe Metin Sınıflandırma başarımına katkısı, Dr. Ahmet Arslan'ın danışmanlığını yürüttüğü bir yüksek lisans tezi kapsamında incelenmektedir. Aynı yüksek lisans tez çalışmasında, bazen bitişik bezen ayrık yazılan birleşik sözcüklerde (örneğin ham petrol, sürçü lisan, gayri menkul, ipek yolu vb.) incelenmektedir.

Elbette yazımda birliğe gidilmesi açısından bunların tespiti kadar doğru biçimde yazılması da önemlidir. Eğer bu yanlışlar otomatik tespit edilebiliyorsa otomatik olarak da düzeltilebilir. Geliştirilebilecek bir bilgisayar programı ile kitap, dergi basım evleri böyle bir düzeltme servisinden faydalanabilirler. Bu yanlışlar Türkçe öğretimi (anadil ya da ikinci dil) sırasında kullanılabilir.

Deneysel sonuçların yeniden üretilebilirliği ve tekrarlanabilirliği (Arguello vd., 2015; Ferro, 2017) hususunu desteklemek amacıyla bu çalışmada kullanılan yazılımın kaynak kodları GitHub https://github.com/iorixxx/Kemik adresinde kamuya açık hale getirilmiştir. Dahası, makaleye sayfa sınırlaması nedeniyle hepsini listeleyemediğimiz yanlış-doğru çiftlerinin tümüne verilen adresten erişilebilir. Böylece diğer araştırmacılar sunulan çalışmadan mümkün olduğunca faydalanabilirler.

KAYNAKÇA

Arguello, J., Crane, M., Diaz, F., vd. (2015). Report on the SIGIR 2015 workshop on reproducibility, inexplicability, and generalizability of results (RIGOR). *ACM SIGIR Forum* 2016; *49*(2): 107–116. http://doi.acm.org/10.1145/2888422.2888439

Arslan, A., Alkılınç, A., ve Dinçer, B.T. (2018). Büyük Veri Setlerinde Varlık Tanıma: En Sık Geçen E-Posta, Web Adreslerinin ve Emojilerin Tespit Edilmesi, *Proceedings of the 6th International Symposium on Innovative Technologies in Engineering and Science*, Antalya, Türkiye (pp. 399–406). https://doi.org/10.33793/acperpro.01.01.79

Białecki, A., Muir, R., ve Ingersoll, G. (2012). Apache Lucene 4. *Proceedings of the SIGIR 2012 workshop on open source information retrieval*, Portland, Oregon, USA (pp. 17–24). http://opensearchlab.otago.ac.nz/paper_10.pdf

Balcı, A. (2017). Sözlüksel Biçimbilim. Genel Dilbilim 1. *Anadolu Üniversitesi AÖF Yayınları*.

Bloomfield, L. (1933). Language. Holt, Reinhart & Winston.

Can, F., Koçberber, S., Balçık, E., Kaynak, C., Öcalan, H. C., ve Vursavaş O. M. (2008). Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, *59*(3), 407–421. https://doi.org/10.1002/asi.20750

Ercilasun, A. B. (2013). Türkçenin dünya dilleri arasındaki yeri. *Dil Araştırmaları*, *12*(12), 17-22.

Ferro N. (2017). Reproducibility challenges in information retrieval evaluation. *Journal of Data and Information Quality*, *8*(2): 8:1–8:4. http://doi.acm.org/10.1145/3020206

Friesner, M. L. (2009). The adaptation of Romanian loanwords from Turkish and French. *Loan phonology*, *307*, 115-129.

Haspelmath, M. (2009). Lexical borrowing: Concepts and issues. *Loanwords in the world's languages: A comparative handbook*, 35-54.

Haspelmath, M., ve Tadmor, U. (Eds.). (2009). *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

Hiemstra, D. (2000). A probabilistic justification for using tf×idf term weighting in information retrieval. International Journal on Digital Libraries, *3*(2), 131-139.

Johanson, L. (1998). Code-copying in Irano-Turkic. *Language Sciences*, *20*(3), 325-337.

Kılınç, D., Özçift, A., Bozyiğit, F., Yıldırım, P., Yücalar, F., ve Borandağ, E. (2017). TTC-3600: A new benchmark dataset for Turkish text categorization. *Journal of Information Science*, *43*(2), 174–185. https://doi.org/10.1177/0165551515620551

Manning, C. D., Raghavan, P., ve Schütze, H. (2008). *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press.

Poplack, S. (2017). *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford University Press.

Sankoff, G. (2002). 25 Linguistic Outcomes of Language Contact. *The handbook of language variation and change*, 638- 668.

Treffers-Daller, J. (2010). Borrowing.http://eprints.uwe.ac.uk/11789/1/output_11789.pdf

Weinreich, U. (1953, yeniden basım: 2010). *Languages in contact: Findings and problems* (No. 1). Walter de Gruyter.

Winter, W. (1960). The Lingua Franca in the Levant: Turkish Nautical Terms of Italian and Greek Origin. *Language*, 36, 3, (1), 454-462.

Yıldırım, O., ve Atık, F. (2013) Yıldız Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Bitirme Projesi.

Zimmer, K. (1985). Arabic loanwords and Turkish phonological structure. *International journal of American linguistics*, *51*(4), 623-625.