TWEETS DERLEMİ
SOSYAL MEDYADAN DERLEM OLUŞTURMAK

Taner SEZER*

Öz: Sosyal medya, çağımızda yaşamın önemli bir parçası haline gelmiştir. Kullanıcılar düşüncelerini, durumlarını, fotoğraflarını, hayatlarındaki dönüm noktalarını ve fikirlerini sosyal medyanın araçlarını kullanarak paylaşmaktadırlar.

Bu çalışmada TS Corpus projesi kapsamında, 1 milyon Tweet ile hazırlanan TweetS derleminin yapım ve kullanım özellikleri incelenecektir. Hazırlanan TweetS derleminde, daha önce Türkçe için sözcük türü etiketlemede hiç kullanılmamış, İnternet diline özgü 6 yeni etiket kullanılmıştır. Veriyi işlemek üzere yeni bir tokenizer (birimlendirici) hazırlanmıştır.

Çalışma, sosyal medyayı kaynak olarak kullanan derlemlere olan ihtiyacı göstermiştir.

Anahtar kelimeler: Derlem, twitter, birimlendirme, sözcük türü işaretleme, ts corpus

* Uzman; Mersin Universitesi, İletişim Fakültesi, Gazetecilik Bölümü.

TWEETS CORPUS; BUILDING A CORPUS BY SOCIAL MEDIA

Taner SEZER*

Abstract

Social media is an important part of the modern life in our era. Users share their ideas, status, photos, life turning points and opinions using instruments of social media.

This study will focus on building and usage features of the TweetS corpus build by using 1 million Tweets, under TS Corpus project. The corpus represents 6 new part-of-speech tags, peculiar to Internet, that were never used before in part-of-speech tagging of Turkish texts before. Also a new tokenizer had prepared in order to process the data.

The study states the need for the corpora that use social media as a data source.

Key words: corpus, twitter, tokenization, part-of-speech tagging, ts corpus

1. Sosyal Medya Çağı

Akıllı telefonlar, tabletler ve dizüstü bilgisayarlar ile taşınabilir hale gelen teknoloji, İnternetin kapsama alanının mobil ağlar ile genişlemesiyle beraber çağımızda kesintisiz olarak çevrim içi kalabilmeyi sağlamaktadır. Bu sürekli çevrim içi olma hali, özellikle "sosyal medya" olarak anılan ağlar ile modern insanın hayatının ayrılmaz bir paçası haline gelmiştir.

Kullanıcılar sosyal medya yoluyla anlık durumlarını, düşüncelerini, yaşamlarındaki belli başlı dönüm noktalarını "*bildirmenin*" yanı sıra, bazen hiç tanımadıkları insanlarla ortak bir zemini paylaşarak gündeme dair fikirlerini de ifade edebilmektedirler.

İnternet ve sosyal medyanın, erişim olanağı olan herkese açık olması, "*belirli bazı konular dışında*" otokontrolün üstüne çıkmayan sansür mekanizması, farklı takma isimler kullanarak gerçek kimliği gizlemeye olanak tanıması ve anlık veya uzun vadeli etkileşime izin vermesiyle mevcut yazım alanlarından farklıdır. Bu açıdan İnternet, içerdiği çok miktardaki veri, metin türlerinin çeşitliliği ve bu verilerin derlem çalışmalarında kullanılabilecek şekilde sayısallaştırılmış olması nedeniyle dilbilim açısından büyük bir doğal veri kaynağıdır (Sharoff, 2006:63).

* Specialist, Mersin University, Faculty of Communication, Department of Journalism.

Çok sayıda kullanıcının, hemen her konuda "*tweet atarak*" katkı sağladığı bu alan, derlem dilbilim çalışmaları için de önemli bir kaynak haline gelmiştir. Scheffler, Twitter'ı "*güncel ve çok çeşitli veri içeren değerli bir kaynak olarak*" tanımlamaktadır (Scheffler, 2014:2284). Bunun temel sebeplerinden biri de, geçmişte sınırlı sayıda kimseye ait bir ayrıcalık olan "*yazma ve yayınlama*" hakkının İnternet ve sosyal medya ile herkese ait bir özgürlük alanı haline gelmiş olmasıdır. Bir "mikro günlük" (micro-blogging) servisi olan Twitter, "*Arap Baharı*" ve "*Occupy*" eylemleri sırasında tüm dünyada ve "*Gezi Olayları*" sırasında da ülkemizde özellikle dikkati çeken bir sosyal medya ağı olmuştur.

Bu çalışmada, 2009-2011 yılları arasında, Türkçe olarak atılan, toplam 1 milyon Tweet'ten (+13 milyon birim) oluşan TweetS Derleminin tasarım aşamalarıyla genel kullanım özellikleri işlenecektir.

2. Sosyal Medyayı İşlemek

İnternet birçok noktada kendine has özellikleri getiren bir teknolojidir. Bunlardan biri de İnternet'e özgü yazım tarzıdır. Yukarıda anıldığı gibi İnternet, erişime sahip herkesin özgürce yazabilmesine olanak sağladığı için, kullanıcılar çok geniş bir sosyokültürel ve sosyoekonomik yelpazeden oluşmaktadır. Kullanıcıların yanı sıra İnternet'in, İnternete erişim için kullanılan cihazların ve erişilen sitenin altyapısı da yazım tarzı üstünde son derece etkilidir.

2.1 Sosyal Medyanın Yazım Tarzı

Twitter'ın "*mikro günlük*" olarak adlandırılmasının temel sebebi, kullanıcılara, her tweet için 140 karakterle sınırlı bir alan sunmasıdır. Bu durum kullanıcıların bazı sözcükleri kısaltarak yazmasını gerektirmektedir. Basılı bir kitapta, gazetede, bilimsel bir yazıda görmenin mümkün olmadığı "*slm cnm, nbr? :)*" şeklinde yazılmış bir cümle İnternet kullanıcıları için son derece sıradan ve kolay anlaşılır bir cümledir.

İnternet dilinde kısaltmalar kullanmak eski bir alışkanlıktır. IRC'nin (*İnternet Relay Chat*) popüler olduğu 2000'li yıllarda "ASL" kısaltması (*age, sex, land - yaş, cinsiyet, ülke*) sohbete hızlı bir başlangıç yapmak için bilinen ve sık kullanılan bir yapıydı. Çevrim içi oyun oynayan oyuncular tarafından kullanılan bir kısaltma ise "AFK". (*away from keyboard*) Bu kısaltmayı kullanıcılar, klavye başında olmadıklarını bildirmek için kullanmaktalar.

Twitter'da gelen verileri işlemek için öncelikle İnternet diline ilişkin yazım unsurlarını ve kısaltmaları işleyebilecek doğal dil işleme araçlarına ihtiyaç vardır. Bu çalışma kapsamında geliştirilen araçlar, 4. bölümde konu edilecektir.

2.2. Teknolojik Sınırlılıklar

İnternete erişimde kullanılan mobil cihazların çoğunda, Türkçe klavye yerleşiminin ya ikincil seviye klavye yerleşiminde yer alması ya da hiç yer almaması, kullanı-

cıların yazım alışkanlıklarını etkileyen bir başka unsurdur. Örneğin mobil cihazlarda, "ş" harfini yazmak için "s" karakterini basılı tutarak ikincil seviyede yazılabilecek karakterleri çağırmak ve bu karakterler içinden "ş'yi" seçmek gerekmekte. Bu durum, hızlı yazmak isteyen kullanıcılar için bir dezavantaj oluşturduğu için bazı kullanıcılar "ş" yerine "s" yazmayı tercih etmektedirler.

Türkçe bir karakteri, Türkçe alfabede bulunmayan bir karakterle değiştirmenin bir diğer sebebi de kullanıcıların tercihlerini bu yönde kullanmaları[1]. Bazı İnternet sitelerinde, örneğin Ekşi Sözlük'te, "ş" karakteri yerine "$" işareti sıklıkla ve bilinçli olarak kullanılmaktadır.

TS TweetS derleminde 545 defa gözlemlenen "*walla*" yazımı bu kullanıma örnek olarak verilebilir.

3. Derlem ve Sosyal Medya

Genel olarak derlemler, dili incelemek veya dilbilimsel çalışma yapmak için kullanılan, büyük ölçekli, sayısallaştırılmış veri setleridir. Sinclair derlemi, "*derlemler, belirli dilbilimsel kriterlere göre seçilmiş ve bir araya getirilmiş, dilin örneği olarak kullanılabilecek metinler bütünü*" olarak tanımlamıştır (Sinclair, 1991:171).

Leech ise derlemleri, "*genellikle belirli bir amaç için bir araya getirilmiş ve bir dili veya metin türünü temsil eden metinlerdir*" (Leech, 1992:105) şeklinde tanımlar.

Bu veri setlerini oluşturmakta kullanılan yöntemler, verilerin toplandığı alanlar ve işleme yöntemleri de teknolojinin ilerlemesiyle birlikte gelişmiştir. Bu kapsamda sosyal medya da derlemler için veri kaynağı olarak kullanılmaya başlanmış ve Twitter'dan elde edilen veriler ile derlemler oluşturulup kullanıcılara sunulmuştur. İngilizce "*Edinburgh Twitter Corpus*" (*ETC*) (Petrovic et al., 2010), Almanca "*A German Twitter Snapshot*" (Scheffler, 2014) bilinen örneklerdir. Türkçe içinse Twitter'dan elde edilen verilerle oluşturulmuş, şu ana kadar yayınlanmış tek örnek "TweetS Corpus" (*http://tscorpus.com/index.php/corpora/tweets-corpus/*) derlemidir.

4. Gerekli Doğal Dil İşleme Araçlarını Üretmek

Doğal dil işleme çalışmalarında (NLP), hedef dile ve amaca uygun araçların geliştirmesini gereklidir. Farklı dillerdeki metinler arasında benzeşmeyen yönler bulunduğu için NLP araçlarının hedef dilin özelliklerine göre tekrar tasarlanması ve yeniden üretilmesi gerekmektedir. Bu araçları "*yeniden üretme gerekliliğine*" sadece dile göre değil, verinin kaynağına göre de ihtiyaç duyulabilir.

Twitter'dan elde edilen veriler için Derczynski, "Twitter metinlerinde sözcük türü işaretleme zordur: veri kirli, dilbilimsel hatalarla dolu ve kendine özgü bir stili var" demektedir (Derczynski et al., 2013:198).

[1] Bu tercihe hangi unsurların sebep olabileceği tartışmaya açık bir konudur.

Twitter verisini incelemek için iki temel noktada kullanmakta olduğumuz mevcut yazılımlarda güncellemeler yapmak yolunu seçtik; ***birimlendirme*** (*tokenization*) ve ***sözcük türü işaretleme*** (*part-of-speech tagging*).

4.1. Birimlendirme (Tokenization)

Birimlendirme, doğal dil işleminin en temel ve önemli adımlarından biridir. Metinleri sözcük türü işaretleme veya biçimbirimsel çözümleme araçlarına, yani daha karmaşık işlemlere göndermeden önce birimlendirmek gereklidir. Birimlendirme işleminde genel olarak kullanılan üç temel yaklaşım vardır. Bunlar ***whitespace tokenization*** (boşlukları temel alarak birimlendirme), ***tree-bank tokenization*** (bir sözcük listesi kullanarak birimlendirme) ve ***RegEx tokenization*** (düzenli ifadelerle birimlendirme) olarak sıralanabilir. Boşluk karakterini temel alarak birimlendirme yapmanın en zayıf noktası, noktalama işaretlerinin önüne ve arkasına birimlendirme işleminden önce birer boşluk karakteri eklenmesidir. Bu yöntem, sosyal medyada sıklıkla kullanılan *smiley* işaretlerinin işlenmesini engellediği için bu çalışmada kullanılabilmesi mümkün değildi.

Benzer şekilde, Derczynski'nin "*kendine özgü yazım stili*" diyerek ifade ettiği İnternet yazımına özgü kısaltmalar ve daha önce anıldığı şekilde, kullanıcıların Türkçe karakter kullanmadan yazma alışkanlıkları (veya mecburiyetleri), bir sözcük listesine bağlı kalarak birimlendirme yapmanın önünde engel olmaktaydı.

Bu çalışmada ürettiğimiz birimlendiricinin aşağıda verdiğimiz temel işlevleri yerine getirebilmesi ve işleyebilmesi gerekmekteydi.

4.1.1. Smiley İşaretleri

Birimlendirici, sosyal medya dilinde son derece önemli olan ve sıklıkla kullanılan ***:)***, ***:D***, ***:p*** gibi smiley işaretlerini işleyebilmeliydi.

4.1.2. Internet'e Özgü Yazım

Birimlendirici, İnternet diline özgü olan, "nbr, slm, cnm, ii" vb kısaltarak yazılan kullanımları işleyebilmeliydi. Bu sözcüklerin "hatalı yazım" olarak adlandırılması ve işlemleme dışında bırakılması, sosyal medyadan veri kullanan bir doğal dil işleme çalışması için son derece verimsiz olacaktır.

4.1.3 İşaret etmek (mention)

Twitter'da kullanıcı adının önünde, boşluk olmadan yazılan "@" işareti, Tweet içinde bir kullanıcıya işaret edildiği anlamına gelmektedir. Örneğin, bir Tweet içinde @foo yazılmışsa, bu foo kullanıcısına işaret edildiğini göstermektedir.

4.1.4. Etiket işareti (hashtag)

Bir Tweet içinde, boşluk olmadan "#" işaretiyle birlikte yazılan sözcükler bir etikete işaret edildiğini göstermektedir.

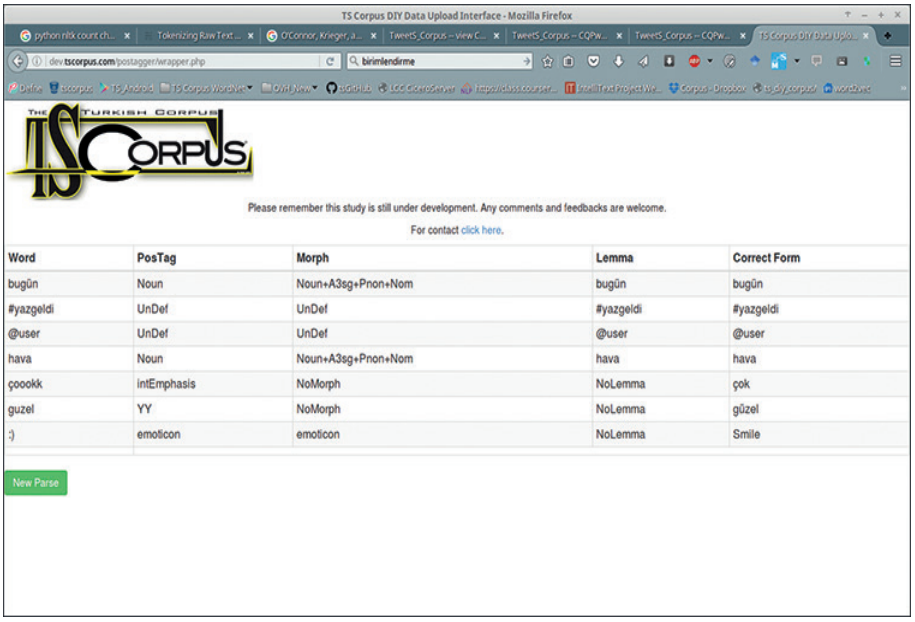
Oluşturduğumuz yeni birimlendirici bu dört özelliği karşılayacak şekilde tasarlandı. Örneğin, "*Bugün #foo hava çok güzel @bar :)*" girdisini birimlendirici "|Bugün||#foo||hava||çok||güzel||@bar||:)| şeklinde ayrıştırabilmektedir.

4.2 Sözcük Türü İşaretleme

Bir metin içindeki sözcüklere, sözcüklerin bulunduğu konumlarındaki işlevlerini belirten etiketlerin iliştirilmesi işi sözcük türü işaretleme (part-of-speech tagging) olarak tanımlanır. Sözcük türü işaretlemede kullanılan etiketler önceden ve belirli ilkeler doğrultusunda belirlenmelidir.

TweetS derlemini etiketlemek için, kullandığımız yazılıma *emoticon*, *intAbbr*, *intSlang*, *intEmphasis*, *YY* ve *intAbbrEng* olmak üzere 6 yeni etiket tanımlayarak, bu etiketlerin veri setinde işaretlemesini yaptık. Ayrıca tanımlanan bu 6 etiketin her biri için, sözcük türü etiketleyicinin ilgili karşılığı göstereceği bir katmanı da çıktı üretirken vermesini ve bu katmanın hem derlem arayüzünde gösterilmesini hem de aramalarda bu etiketlerin kullanılmasını sağladık. Aşağıda verilen ekran görüntüsünde, sözcük türü işaretleyicinin ürettiği örnek bir çıktı görülmektedir.

Resim 1. Hazırladığımız birimlendirici[2] ve sözcük türü işaretleyicinin[3] ortaklaşa ürettikleri, "bugün #yazgeldi #user hava çoookk güzel :)" tweet'inin işlenmiş hali.



Word	PosTag	Morph	Lemma	Correct Form
bugün	Noun	Noun+A3sg+Pnon+Nom	bugün	bugün
#yazgeldi	UnDef	UnDef	#yazgeldi	#yazgeldi
@user	UnDef	UnDef	@user	@user
hava	Noun	Noun+A3sg+Pnon+Nom	hava	hava
çoookk	intEmphasis	NoMorph	NoLemma	çok
guzel	YY	NoMorph	NoLemma	güzel
:)	emoticon	emoticon	NoLemma	Smile

[2] TS Tokenizer çevrim içi kullanılabilir bir yapıda kullanıcı erişimine açıktır: http://dev.tscorpus.com/tokenizer/index.php

[3] Hazırladığımız sözcük türü işaretleyici çevrim içi olarak kullanılabilir şekilde kullanıcı erişimine açıktır: http://dev.tscorpus.com/postagger/index.php

Bu noktada derlem dilbilim için sözcük (*word*), birim (*token*) ve kök (*lemma*) kavramlarını açıklamak, yukarıdaki çıktının anlaşılması için verimli olacaktır.

Evert (2005:18) sözcüğü, "*teorinin tanımına veya hedeflenen uygulamaya göre, herhangi bir sözlüksel öğeye işaret edebilecek son derece genel bir ifade*[4]" olarak tanımlamaktadır. Bu tanımın birim (token) için de geçerli olduğunu söyleyebiliriz.

TweetS Derlemini işaretlerken temel aldığımız kriterlerde ***sözcük***, veri kaynağında görüldüğü halde derleme (*ve derlem arayüzüne*) aktarılan görünümdür. ***Birim***, sözcükleri ve bununla birlikte noktalama işaretleri, emoticon ifadelerini, sayıları, kısaltmaları vb. sözcük olmayan, bir noktada sözlükte yer almayan, ancak veri içinde bulunan tüm öğeleri kapsar. ***Kök*** ise, sözcüğün işaretleme sırasında bulunan olası en ham halini ifade eder.

4.2.1. intAbbr

Bu etiket altında İnternet kullanıcılarının, sıklıkla kullandığı, kısaltarak yazılan sözcükler yer almakta. Kullanıcıların "iyi" yerine "ii", "değil" yerine "deil" yazmayı tercih ettiği yazımlar bu etiketle işaretlenmiştir. Bu yazım tarzının, yukarıda anıldığı şekilde, sosyal medya ve İnternet dili için, yazım hatasından farklı değerlendirilmesi gerekmektedir.

4.2.2. emoticon

İngilizce "emotion" (*duygu*) ve "icon" (*ikon*) sözcüklerinden üretilen "emoticon" sözcüğü, noktalama işaretleriyle yüz ve beden ifadelerini göstermek için kullanılan ikonlara verilen genel addır. Bu ifadeler İnternet'te sıklıkla kullanılmaktadır. Derlemin, duygu analizi (*sentiment analysis*) konusunda çalışmak isteyen araştırmacılar için de verimli bir kaynak olması için bu ifadelerin etiketlenmesinin yararlı olacağına düşündük.

Bunun ötesinde, Türkçe bir derlemde, ilk defa emoticon ifadelerinin etiketlenmesini yaparak, bu ifadeler için tanımlanan etiket ile arama yapmasına olanak sağlamış olduk.

Bu emoticon simgelerinin karşılık geldikleri ifadeler derlem arayüzünde gösterilmektedir.

4.2.3. intEmphasis

Yazılı dilin sözlü dilden temel farklarından biri de tonlama ile belirli noktalara vurgu yapılamamasıdır. Ancak, İnternet gibi, dilin yazım kurallarına uymanın gerekli olmadığı alanlarda, zaman içinde kullanıcılar bu vurguları yapmanın çeşitli yöntemlerini geliştirdikleri gözlenmiştir. Örneğin "çok güzeldi çoookk" yazımında, ikinci *çok* sözcüğü vurgulama amacıyla, bilinçli bir tercih sonucunda "çoookk" şeklinde yazıl-

[4] "…. entirely generic term which may refer to any kind of lexical item, depending on the underlying theory or intended application."

mıştır. Bu sözcüklerin, kurallı yazımda olması beklenilen yazımları derlem arayüzünde gösterilmektedir.

4.2.4. intAbbrEng

İnternette kullanılan bazı kısaltmalar İngilizce sözcüklerden gelmekte olup, Twitter hayata geçmeden önce de kullanılmaktaydı. Bu kısaltmaları ayrı bir kategoride topladık.

4.2.5. intSlang

İnternet argonun bolca kullanıldığı bir mecra. Argo sözcükleri kendine ait bir kategoride topladık.

4.2.6. YY

Bu kategori altında, *intAbbr* etiketindeki sözcüklerden farklı olarak, Türkçe karakterleri kullanmadan yazılan, *degilim*, *tesekkurler*, *guzel* gibi sözcükler yer almakta. Bu sözcükler için ayrı bir kategori oluşturmanın, İnternet dilinin araştırılmasına katkıda sağlayacağını düşünerek, farklı bir kategori altında topladık. "Yazım yanlışı" olarak etiketlenen sözcüklerin doğru yazımı derlem arayüzünde gösterilmektedir.

5. Derlemin Altyapısı ve Kullanımı

5.1. Derlem Altyapısı

TweetS Derlemi (*ve TS Corpus projesi altında yer alan tüm diğer derlemler*) CQPWeb/CWB[5] altyapısını kullanmaktadır. Bu yapı, üst metin ve dilbilimsel işaretlemeleriyle birlikte 2 milyar sözcüğe kadar büyüklükteki verinin web arayüzü ile kullanıcıların erişimine sunulmasına izin vermektedir.

TS Corpus projesi Fransa'da, OVH veri merkezinde, "*dedicated*" bir sunucuda barındırılmaktadır. Derlemler ve projenin sunduğu diğer NLP servisleri, Debian 8.2 x86_64 işletim sistemi üstünde çalışmaktadır. Donanım olarak Intel® Xeon® D-1520 işlemci ve 32 Gb RAM ile desteklenen sistem, 1 Gbps bağlantı hızına sahiptir.

5.1.1 Derleme Erişim

Kullanıcıların derleme erişmek için projenin ana sayfası olan http://tscorpus.com adresinden üye olmaları gerekmektedir. Üyelik, kullanılan versiyonda e-posta yoluyla aktive edilmektedir. Kullanıcılar, sisteme üye olduklarında tanımlanmış tüm derlemlere önceden belirlenmiş yetkiler çerçevesinde erişebilmektedirler.

Kullanıcı sisteme giriş yaptığında karşılama ekranına yönlendirilir. Bu ekran, aktif olan tüm derlemleri listelemektedir. TweetS Corpus bağlantısına tıklayarak kullanıcılar derlem sorgu arayüzüne erişebilirler. Bu ekran kullanıcıların erişebilecekleri diğer özellikleri içeren bir menüyü ve arama anahtarını girecekleri metin kutucuğunu içeren, kullanıcı dostu bir arayüz ile tasarlanmıştır.

[5] http://cwb.sourceforge.net/

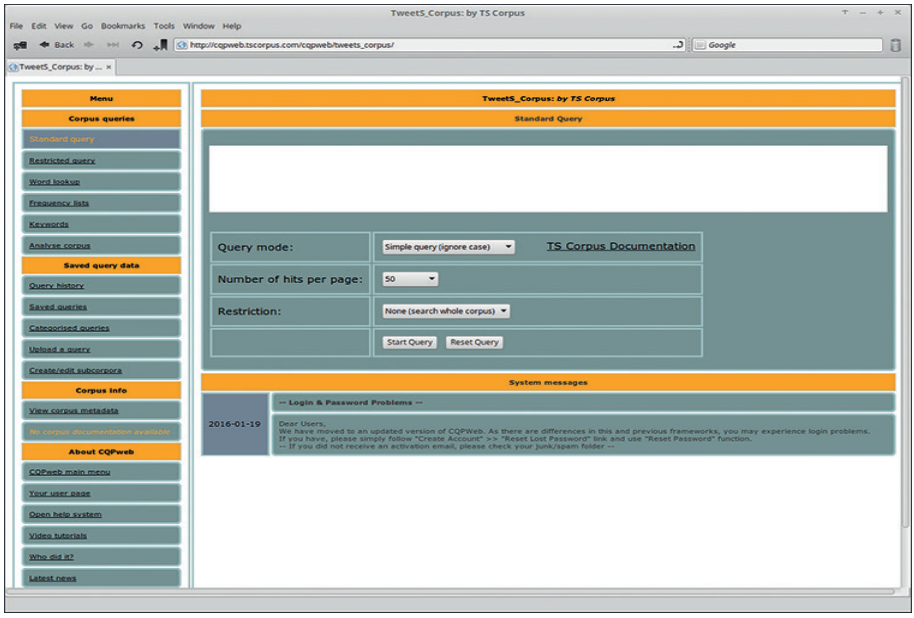
5.1.2. Kullanıcı Arayüzü

Derleme erişildiğinde, ekranın solunda görülen menüler ile kullanıcılar "***sözcük arama***", "***sıklık listeleri***", "***anahtar sözcükler***" gibi fonksiyonlara erişebilir, kendi kullanıcılarına ilişkin tercihlerini yaparak bunları saklayabilirler. CQPWeb/CWB kullanıcıların alt derlemler oluşturmasına, arama sonuçlarını saklamasına ve arama sonuçlarına ilişkin kategoriler belirlemesine ve bu kategorileri saklamalarına izin vermektedir.

"***Query Mode***" seçeneği ile kullanıcılar sorgularının büyük/küçük harf duyarlı veya duyarsız olmasını ayarlayabilir veya sorgularında CQP[6] dilini kullanmayı seçebilirler. Sorgu, arama kutucuğuna bir anahtar yazıp "***Start Query***" düğmesini tıklayarak başlatılır. Örnek bir arama sonucu aşağıdaki ekran görüntüsünde verilmiştir.

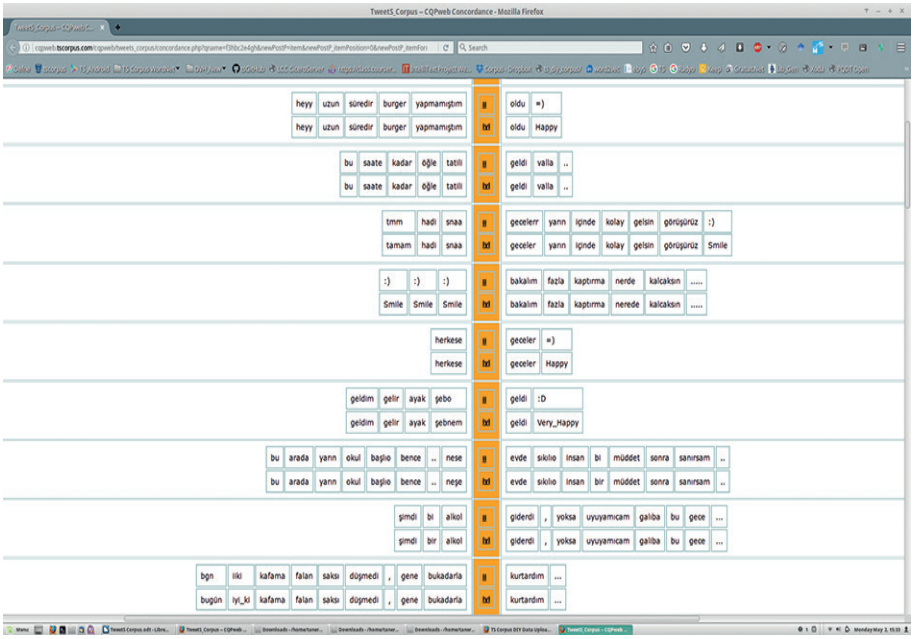
Arayüzün mobil cihazlara uyumlu hale getirilmesi amacıyla çalışmalar sürmektedir.

Resim 2. TweetS Derlemi ana sorgu ekranı



[6] Corpus Query Processor

*Resim 3. *_emoticon sözcük türü etiketi için döndürülen sonuçlar.*

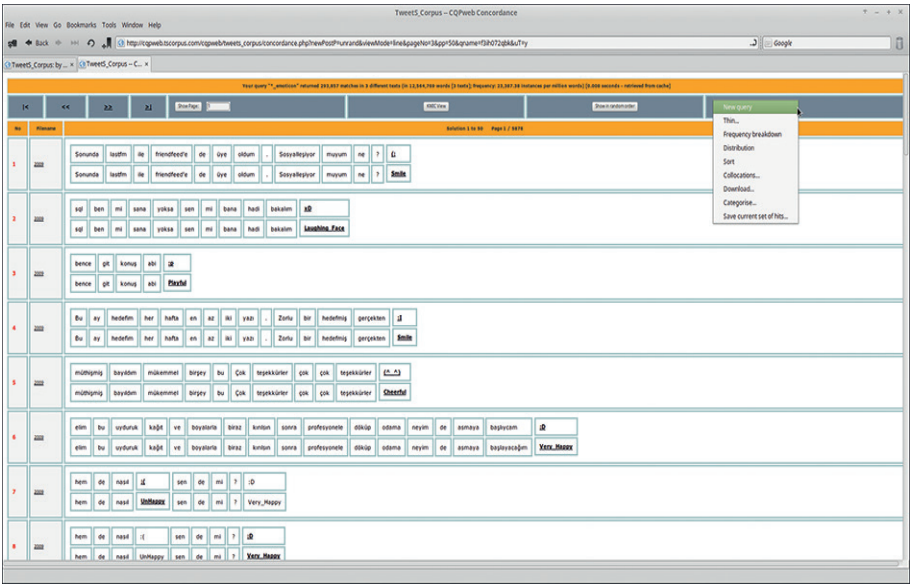


Sonuçlar iki satır halinde gösterilmektedir. İkinci satırda emoticon ifadelerinin açıklamaları, kısaltmaların açılımları, YY (*yazım yanlışı*) etiketiyle işaretlenen sözcüklerin doğru yazım önerileri bulunmaktadır. TweetS Derlemi öntanımlı olarak KWIC[7] (bağlam içinde anahtar sözcük) görünümünde sonuçları getirecektir. Ekranın en üst satırında bulunan "***Line View***" (satır görünümü) düğmesi ile bu görünüm değiştirilebilir. Aşağıdaki ekran görüntüsünde satır görünümü örneklendirilmiştir. Aynı görüntüde, ekranın sağ üst köşesinde bulunan menü "*action menüsü*" olarak adlandırılır. Bu menü altındaki seçenekler ile bulunan sonuçlar üstünde işlem yapmak mümkündür.

Bu menü, ***Thin*** (sonuçları azaltma), ***Frequency Breakdown*** (sıklık dağılımı), ***Distribution*** (dağılım), ***Sort*** (sıralama), ***Collocations*** (eşdizim örüntüleri), ***Download*** (sonuçları yerel diske indirme), ***Categorise*** (sınıflandırma) ve ***Save Current Set of Hits*** (aktif sonuçları kaydetme) seçeneklerini barındırır.

[7] Keyword In Context

Resim 4. Action menüsü ve satır görünümü



5.2. Temel Sorgular

Derlem arayüzünde iki tarz sorgu oluşturmak mümkündür; ***basit aramalar*** (*basic queries*) ve ***CQP dilinde*** (*CQP Syntax*) oluşturulan sorgular. Her iki sorgu türünün de kendine özgü farklılıkları, avantajları ve yazım dili (*syntax*) vardır. Kullanıcılar, hedefleri doğrultusunda bu sorgu türlerinden birini seçebilirler.

5.2.1. Basit Sorgular (Basic Queries)

Basit sorgular sadece anahtar kelimeyi girerek, bu sorguyu (*düzenli ifadelerde de kullanılan*) joker karakterler ile destekleyerek veya sözcük türü etiketlerinden birini kullanarak yapılır. Basit bir aramayı başlatmak için anahtar sözcüğü arama kutucuğuna yazıp "Start Query" düğmesini tıklamak yeterli olacaktır. Örneğin "*yeni*" anahtar sözcüğü derlem içindeki tüm "*yeni*" sözcüklerini getirecektir. Eğer sorgu anahtarı "yeni*" şeklinde verilirse "yeni" dizimini içeren ve diziyi takip eden tüm karakterleri barındıran sözcükler, örneğin *yenilik*, *yenice* vb. sonuç olarak getirilecektir.

Basit aramalarda köşeli parantez ([ ]) değişken karakterleri tanımlamak için kullanılabilir. Sorgu anahtarı olarak t[o,ü]p girilirse, t ve p harfleri arasında sadece o ve ü harflerini içeren sözcükler sonuç olarak *gelecek*, *tip*, *tep* gibi sözcükler işlem dışında tutulacaktır.

Bir sözcüğü (*istenirse düzenli ifadelerle de destekleyerek*) sözcük türü etiketiyle birlikte sorgu anahtarı olarak kullanmak da mümkündür. gel*_Noun sorgusu, *gel* dizisi dahil olmak üzere, bu dizilimle başlayan ve herhangi bir sayıda karakter dizisini içeren ve sözcük türü Noun (isim) olarak etiketlenmiş tüm sonuçları listeleyecektir.

Basit aramalar aynı zamanda ***lemma*** (*kök*) sorguları için de kullanılabilir. Bunun için sorgu anahtarı süslü parantez (*curly braces*) içinde yazılmalıdır. Örneğin, sorgu anahtarı {burun} şeklinde verilirse, *burun* sözcüğü de dahil olmak üzere bu kökü içeren tüm görünümler listelenecektir; *burnu*, *burnum*, *burnunda* vb.. Özellikle ses düşmesi vb. ses olaylarının gözlemlendiği sözcüklerde lemma sorguları verimli bir kullanım sağlamaktadır.

5.2.2 Gelişmiş Sorgular (CQP Syntax Queries)

CQP'nin kendi özgü bir dili vardır. Bu dil, ancak CQP'nin Linux kabuğu (bash) üstünde kullanılan ara biriminde tam olarak çalışsa da, CQPWeb bazı özelliklerin web arayüzü üstünden kullanılmasına da izin vermektedir. Bu sorgular genellikle karmaşık ve dilbilimsel işaretlemenin bulunduğu seviyede kullanılır.

Örneğin, *görmek* eyleminin geçmiş zaman eki alarak kullanıldığı bağlamı sorgulamak için CQP Syntax modunda [Lemma="gör" & Morph=".*\+Past\+.*"] sorgu anahtarı girilmelidir. Bu anahtar, Lemma içinde verilen kök üstünde biçimbirimsel etiket olarak "*Past*" kullanılan tüm yapıları getirecektir.

Bir başka kullanım örneği olarak şu sorguyu verebiliriz:

[Lemma="televizyon"] [Lemma="izle" & PosTag="Verb" | Lemma="seyret" & PosTag="Verb" | Lemma="bak" & PosTag="Verb"]

Bu sorgu televizyon sözcüğünden hemen sonra gelen ve sözcük türü Verb (*eylem*) olarak işaretlenmiş tüm yapıları tek bir sonuç listesi halinde getirecektir. Böylelikle kullanımlar arasındaki ilişkiler veya farkılıklar kolaylıkla gözlenebilir.

6. Sonuçlar

TweetS derlemi, Tweet'ler ile oluşturulmuş, çevrim içi olarak erişilebilen, sözcük türü ve biçimbirimsel olarak etiketlenmiş ilk Türkçe derlem çalışmasıdır. Derlem, yayınlandığı tarihten bu yana bilimsel çalışmalarda ve çeşitli doğal dil araştırmalarında kaynak olarak kullanılmış olup TweetS derlemini veri seti olarak kullanan bir yüksek lisans tezi yazılmıştır Karatay, TS Corpus'u, "*Türkçe'de sıklıkla kullanılan ifadelerin istatistiğini veren, ihtiyaçları karşılayan ve alandaki boşluğu dolduran bir derlem*" olarak tanımlamıştır (Karatay, 2014:25). Kullanıcılar, TweetS derleminde bugüne kadar 120 binden fazla sorgu yapmıştır.

Bu göstergeler, sosyal medyayı veri seti olarak kullanan derlemlerin önemini vurgulamaktadır. Bu bağlamda daha fazla ve daha güncel veri içeren yeni derlemlere ihtiyaç vardır.

TweetS derlemini hazırlarken, veriyi işlemek amacıyla oluşturduğumuz birimlendirici, bu veri setine özgü araçların gerekliliğini vurgulamaktadır.

Sözcük türü etiketleyiciye tanımladığımız yeni etiketler, "*sosyal medyanın kendine özgü dilini*" daha verimli işlemek üzere eldeki mevcut yazılımların ve etiket setlerinin geliştirilebileceğini göstermektedir.

Sosyal medya bağlamını, Web 2.0 teknolojisinin getirdiği "*etkileşim olanakları*" olarak genişleterek, çok daha fazla kullanıcının dil üretimini içeren derlemlerin oluşturulması, Türkçe derlem dilbilim çalışmalarına katkı sağlayacaktır.

Kaynakça

Cöltekin, C. (2010, May). A Freely Available Morphological Analyzer for Turkish. In *LREC*.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013, September). Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *RANLP* (pp. 198-206).

Evert, S. (2005). The statistics of word cooccurrences (Doctoral dissertation, Dissertation, Stuttgart University)

Karatay, D. (2014). *Tweet Recommentdation Under User Interest Modeling With Named Entity Recognotion* (Doctoral dissertation, Middle East Technical University).

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Startvik (Ed.), *Directions in corpus linguistics* (pp. 105-122). Berlin: Mouton de Gruyter.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Petrovic, S., Osborne, M., & Lavrenko, V. (2010, June). The Edinburgh Twitter Corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media* (pp. 25-26).

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. In *Advances in natural language processing* (pp. 417-427). Springer Berlin Heidelberg.

Scheffler, T. (2014). A German Twitter Snapshot. In *LREC* (pp. 2284-2289).

Sezer, B., Sezer, T. 2013. *TS Corpus: Herkes için Türkçe Derlem.* 27. Ulusal Dilbilim Kurultayı Bildiri Kitabı. 3-4 Mayıs 2013. Antalya, Kemer: Hacettepe Üniversitesi, İngiliz Dilbilim Bölümü, 217-225

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. *WaCky*, 63-98.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.