SENTETİK TÜRKÇE SÖZCÜK KÖKLERİ ÜRETİMİ

Gülşen Cebiroğlu[1] A.Cüneyd Tantuğ[2] Eşref Adalı[3] Yaşar Erenler[4]

gulsen@cs.itu.edu.tr tantug@cs.itu.edu.tr adali@cs.itu.edu.tr erenler@cs.itu.edu.tr
İstanbul Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği, 34390, Maslak, İstanbul, Türkiye

Anahtar Kelimeler: Doğal Dil İşleme, Sentetik Sözcük Kökü Üretimi

ÖZET

Tükçe, bütün dünya dilleri arasında kurallı olması ve düzenli yapısıyla farklı bir konumdadır. Bu çalışma kapsamında Türkçe'nin dilbilgisi kuralları kullanılarak geliştirilen bir bilgisayar yazılımı ile olası yeni sözcük kökleri üretilmiştir. Gerçeklenen çalışmada PC-Kimmo yazılımı için geliştirilmiş olan Türkçe sözlükler kullanılarak yaklaşık 24 bin sözcüklü bir sözcük bankası oluşturulmuş ve üretilen sözcüklerin Türkçe'de kullanılıp kullanılmadığı, bu sözcük bankası ve üretilen sözcük kökleri arasında yapılan istatistiksel karşılaştırmalarla ortaya çıkarılmaya çalışılmıştır.

I.GİRİŞ

Doğal dil işleme alanında, Türkçe'nin farklı bir konumda olmasının temel nedeni, Türkçe'nin dil bilgisi yapısının kurallara dayanması, düzenli olması ve istisnaların sayısının az olmasıdır. Bu özellikleri ile Türkçe, bilgisayar ile işlenmek ve analiz edilmek için çok uygundur.

Türkçe konusunda çalışan dil bilimcilerinin en önemli sorunlarından bir tanesi, yeni ortaya çıkan kavramların Türkçe karşılıklarının türetilmesidir. Çoğu kez, teknolojik ya da kültürel alanda ortaya çıkan yeni kavramlara, zamanında Türkçe karşılık bulunamadığı için, bu kavramları ifade eden yabancı sözcükler zamanla halk tarafından kabul görmekte ve bir daha değiştirilememek üzere Türkçe'ye yerleşmektedir. Bu sorunun tek çözümü, yeni kavramlara karşılık gelen Türkçe sözcüklerin zamanında üretilmesi veya türetilmesidir [1].

Bu çalışmada, Türkçe'deki sözcüklerin özellikleri incelenmiş, harflerin yanyana gelme durumları ve sözcük içerisinde bulunabilecekleri yerlere ilişkin kurallar araştırılmış, bulunan kurallar kullanılarak Türkçe harfleri bitiştirme yoluyla olası sözcük kökleri üreten bir yazılım gerçeklenmiştir. Günlük hayatta kullanılmakta olan Türkçe kökenli sözcüklerin ses kurallarına uyanlarının büyük bir çoğunluğu, sentetik olarak üretilen bu sözcük köklerinin içerisinde yer almaktadır. Ancak üretilen sözcük köklerininin çoğunluğu, henüz bir anlam taşımayan, üzerine anlam yüklenebilecek, Türkçe kurallara uygun sözcüklerdir. Çalışmamızdaki ana amaç, olası sözcük köklerini dil bilimcilere sunarak, yeni kavramlara Türkçe karşılıklar üretilmesini kolaylaştırmaktır.

Çalışma üç temel bölüme ayrılmıştır. Birinci bölümde, Türkçe sözcükleri oluşturan harflerin diziliş kuralları ortaya konulmuş ve bu kurallar kullanılarak harfleri bitiştirip yeni Türkçe sözcük kökleri üreten bir yazılım gerçeklenmiştir. Üretilen sözcük köklerinin kullanılabilirliğini ölçmek ve günümüz Türkçesi'ndeki sözcüklerin ne kadarının bu kurallara uyduğunu belirlemek üzere bir Türkçe sözcük bankası oluşturulması süreci çalışmanın ikinci aşamasını oluşturmaktadır. Son bölümde ise sentetik olarak üretilen sözcüklerin, Türkçe sözcük bankasında bulunma oranları, sözcüklerin harf sayılarına göre sözcük bankasında bulunma oranları, sözcük boylarının dağılımları ve benzeri sonuçlar incelenmiştir.

II.SENTETİK SÖZCÜK ÜRETİMİ

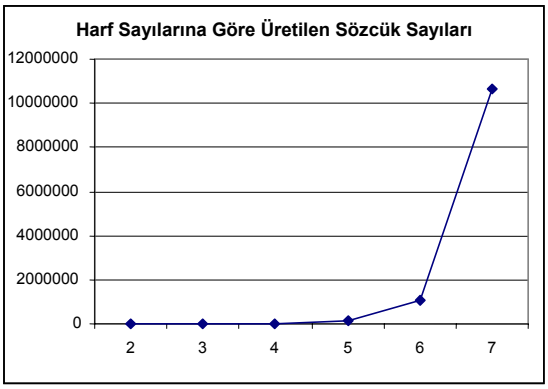
Bilgisayar yardımı ile Türkçe sözcük üretimine başlamadan önce, Türkçe sözcüklerdeki harflerin diziliş kurallarının araştırılması ve bu kuralların hangilerinin uygulanacağına karar vermek gereklidir. Türkçe sözcüklerdeki harflerin diziliş sırası ve harflerin sözcük içinde bulundukları yerlere ilişkin kurallar çeşitli kaynaklarda [1,2,3,4] aşağıdaki gibi listelenmiştir:

- Büyük sesli uyumu kuralı
- Küçük sesli uyumu kuralı
- Sessiz uyumu kuralı
- Sözcük sonunda bulunan çift sessiz harf kuralı
- Sözcük başında c, f, ğ, h, j, l, m, n, p, r, ş, v, z bulunmaması kuralı
- Sözcük sonunda b,c,d,g harflerinin bulunmaması kuralı
- Sözcük başında 2 sessiz harf bulunmaması kuralı
- Sözcük sonunda 3 sessiz harf bulunmaması kuralı
- n ve b harflerinin yanyana gelmemesi kuralı

Geliştirilen yazılım dahilinde sözcük üretiminde yukarıda anılan kuralların dışında çeşitli sınırlamalar da getirilmiştir. Bunlardan ilki, üretilecek olan sözcük köklerinin uzunluğudur. Yazılımın ürettiği sözcüklerin uzunluğu en fazla kullanıcının belirlediği harf adedi kadar olabilir. Ayrıca geliştirilen yazılıma başlangıç kriterleri verilebilme özelliği de getirilmiştir. Üretilecek sözcüklerin başlangıç harfi ya da harfleri, kullanıcı tarafından belirtilebilmektedir. Bu özellik sayesinde kullanıcılara, istedikleri harf veya harf kümesi ile başlayan yeni sözcükler üretebilme olanağı getirilmiştir. Tablo 1'de sentetik sözcük üretme yazılımı ile üretmiş olduğumuz sözcük köklerinin sayıları, Şekil 3'de ise bu sayıların grafiği verilmiştir.

Tablo 1: Üretilen Sentetik Sözcük Kökü Sayıları

Harf Adedi	Üretilen Sözcük Kökü Sayısı
2	240
3	1416
4	12512
5	146416
6	1058512
7	10618784
TOPLAM	**11837848**



Şekil 3: Üretilen Sözcük Köklerinin Harf Sayılarına Göre Dağılımları

Tablo 1'de de görüldüğü gibi en uzun 7 harfli sözcükler oluşturulmuştur. Üretilenlerin, olası sözcük kökleri olduğu, bu köklerin daha sonra yapılacak çalışmalarla yapım ekleri kullanılarak daha farklı sözcükler olarak türetilebileceği göz önüne alındığında bu uzunluk yeterli görülmüştür. Ayrıca 8 harfli sentetik sözcük kökleri, kabul edilebilir süreler ve kaynaklar kullanılarak üretilememiştir. 8 harfli sözcük kökleri üretilse dahi, sayısı çok olacağından, 3. aşamada gerçeklenen karşılaştırmalar için uygun bir zemin oluşturmayacaktır.

Harf sayısının az olmasına karşın üretilen sözcük sayısının fazla olması, çalışmanın amacı olan yeni kavramlara Türkçe karşılıkların, üretilen sözcük kökleri arasından bulunması işlemini zorlaştırmaktadır. Sonuç kümesini daraltmak için, geliştirilen yazılıma uygun başlangıç harf(ler)i verilerek daha az sayıda sözcük üretilebilir. Bu sayede daha gerçekçi bir sözcük uzayında arama yapılarak olası en iyi karşılık aranabilmektedir. Örneğin "bilgi" ile başlayan en fazla 8 harfli olası sözcük kökleri üretilirse toplam 843 adet kök üretilmektedir. Üretilen kökler arasında "bilgibeç", "bilgibel", "bilgiz", "bilgizli" gibi kullanılabilirliği yüksek olan sözcük kökleri bulunmaktadır. Örnek olarak "al" ile başlayan en uzunu 5 harfli Türkçe sözcüklerin üretilmesi sonucunda, bir kısmı aşağıda verilen toplam 780 adet sözcük oluşturulmuştur:

alaca	alaka	alanç	alara
alaş	alçı	alçıt	alaz
alay	albay	alçak	alçar
algan	alfat	algıt	alkıt
algıç	algın	alıt	alhan
alkış	alkat	allaz	allıç
alman	almız	alpak	alpar
alpıt	alraç	alsay	alşıl
altın	alvaz	alzat	alzak
alzar	alzıt	alzır	altaç

Verilen örnek sözcüklerde de görüldüğü üzere, üretilenlerden bazıları halen kullanılmakta olan, diğerleri ise üzerine yeni anlamlar yüklenebilecek, Türkçe kurallarına uygun köklerdir. Türkçe'ye yeni bir kavram eklenirken en sık yapılan hatalardan biri, İngilizce bir kavramın Fransızca okunuşu ile dilimize yerleştirilmesidir. Yabancı kökenli "televizyon" ve "radyo" sözcüklerinin İngilizce okunuşları "televijın" ve "redio" iken bu sözcükler dilimize Fransızca okunuşları olan "televizyon" ve "radyo" olarak yerleşmiştir. Türkçe için de, Almanya, Fransa, Macaristan gibi ülkelerin dillerini yabancı sözcüklerden korumak için başlattığı, "dil gümrüğü" adı verilebilecek bir uygulamaya gidilmelidir [1]. Çalışmamız, dil bilimcilerine kolaylık sağlamayı hedeflemektedir. Örneğin "reseptör" terimi Türkçe'ye yerleşmeden önce dil gümrüğünde karşılığı bulunsaydı, ürettiğimiz sözcüklerden "algıç" veya "algıt" sözcükleri kullanılabilirdi.

Üretilen sözcük köklerinin sayısının çok olması, yeni kavramlar için bulunacak yeni sözcüklerin seçiminde alternatiflerin çok olması anlamına gelir ki bu da bize Türkçe'nin sözcük üretimi açısından potansiyeli yüksek bir dil olduğunu göstermektedir.

III.SÖZCÜK BANKASININ OLUŞTURULMASI

Sentetik olarak üretilen sözcüklerin ne kadarının günümüz Türkçesinde kullanıldığının, günümüz Türkçesinde yer alan sözcüklerinin de kaçının üretilen sözcükler arasında yer aldığının yani Türkçe ses kurallarına uyduğunun araştırılması amaçlanmış ve bu doğrultuda Türkçe sözcüklerden oluşmuş bir sözcük bankası oluşturulmuştur. Elektronik ortamda tutulan bir Türkçe sözlüğe erişim sağlanamadığından yeni bir sözcük bankası oluşturulmuştur. Bunun için, PC Kimmo yazılımı için hazırlanmış Türkçe Sözlük (Turklex) [5] ve Türk Dil Kurumu'nun İmla Kılavuzu [6] kullanılmıştır. Bu kaynaklardaki sözcükler kullanılarak bir sözcük bankası oluşturulmuştur. Üretilen sözcüklerin harf bazında karşılaştırılması yapılacağı için, oluşturulan sözcük bankası içindeki sözcükler, geliştirilen bir yazılım ile harf sayılarına göre ayrıştırılmıştır. Sözcük bankasındaki sözcüklerin harf sayılarına göre dağılımları Tablo 2'de ve Şekil 4'de görülmektedir.

Tablo 2: Sözcük Bankasının Harf Sayılarına Göre Dağılımı

Harf Sayısı	Sözcük Sayısı	Harf Sayısı	Sözcük Sayısı
2	119	12	1419
3	920	13	851
4	2331	14	512
5	6021	15	231
6	6080	16	129
7	7218	17	44
8	7286	18	24
9	5270	19	9
10	4614	20	2
11	2979	TOPLAM	22689

Şekil 4: Sözcük Bankasındaki Sözcüklerin Harf Sayılarına Göre Dağılım Grafiği

Sözcük bankasındaki dağılımın grafiği incelendiğinde, 7-8 harfe sahip sözcüklerin sayısının tepe değerler olduğu görülmektedir. Bu inceleme sonucunda 7 harften daha uzun sözcüklerin çoğunlukla mastar eki (-mek, -mak) almış olan eylemler ve birleşik isimlerden oluştuğu görülmüştür. Bu nedenle uzunluğu en fazla 7 harfli sözcükler üretilmiştir. Üretilen sözcükler Türkçe kurallarına uygun olarak üretildikleri için, günümüz Türkçe'sindeki Türkçe kökenli sözcüklerin sayısının bilinmesi de önemlidir. Türk Dil Kurumu'nun verilerine göre, TDK Türkçe Sözlüğü'nde yer alan sözcüklerin dağılımı Tablo 3'deki gibidir.

Tablo 3: TDK Türkçe Sözlüğü'ndeki (1998) Sözcüklerin Kökenleri

Diğer Diller	14.394
Türkçe	46.301
TOPLAM	60.695

IV.KARŞILAŞTIRMALAR

Gerçeklenen sentetik sözcük kökü üretim yazılımı ile üretilen, en uzunu 7 harften oluşan sözcük köklerinin kullanılma oranlarını araştırmak, günümüz Türkçesinde kullanılan sözcüklerden ne kadarının üretilebildiğini ölçmek ve sözcüklerin ne kadarının Türkçe ses kurallarına uyduğunu incelemek için üretilen sentetik sözcük kökleri ve oluşturulan sözcük bankası arasında bazı karşılaştırmalar yapılmıştır. İlk olarak, üretilen sözcük köklerinin sözcük bankasında bulunması durumu incelenmiştir. Bu incelemenin harf sayılarına göre sayısal sonuçları Tablo 4'de, grafik sonuçları da Şekil 5'de verilmiştir.

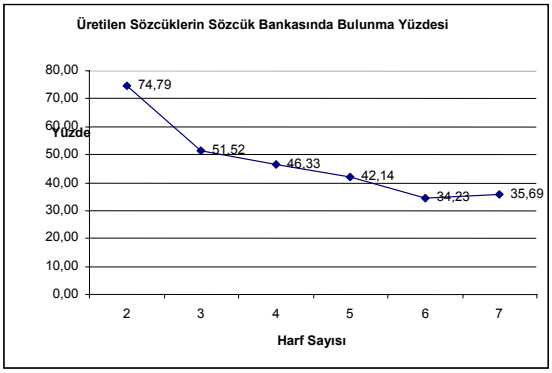
Tablo 4: Sözcük Bankasında Bulunan Sentetik Sözcük Kökleri Adetleri ve Yüzdeleri

Harf Sayısı	Sözcük Bankasında Bulunan	
	(Sayı)	(Yüzde)
2	89	%74.8
3	474	%51.5
4	1080	%46.3
5	2537	%42.1
6	2081	%34,23
7	2576	%35.7
	TOPLAM: 8837	**ORTALAMA : %38,95**

Tablo 4'den ve Şekil 5'de yer alan grafikten görülebileceği gibi harf sayısı arttıkça üretilen sözcüklerin, sözcük bankasında bulunma oranı düşmektedir. Sözcük bankasındaki tüm sözcüklerin üretilememesinin nedenleri arasında kullanılan sözcük bankasında yer alan sözcüklerin tamamının Türkçe kökenli olmaması, birleşik isimlerin ve kuraldışı sözcüklerin bulunması sayılabilir.

Gerçeklenen bir diğer analiz de üretilen sözcüklerin ne kadarının sözlükte bulunduğudur. Bu analizde elde edilen sonuçlar da olası Türkçe sözcük köklerinin ne kadarlık bir bölümünün günümüz Türkçe'sinde kullanıldığını ortaya çıkarmaktadır. Tablo 5'de, üretilen sözcüklerin ne kadarlık bir yüzdesinin sözcük bankası içerisinde yer aldığı görülmektedir.

Tablodan da görülebileceği gibi, harf sayısı arttıkça, üretilen olası sözcük kökü sayısının aşırı miktarda artmasından dolayı, üretilen sözcüklerin büyük bir bölümü sözcük bankasında bulunmamaktadır. Bu ise bu sözcüklerin henüz Türkçe'de kullanılmadığını ve yeni kavramları ifade etmek için bu sözcüklerden yararlanılabileceğini göstermektedir. Bu oranların düşük olması bir anlamda Türkçe'nin sözcük türetme potansiyelinin yüksek olduğunu kanıtlamaktadır.



Şekil 5: Üretilen Sentetik Sözcük Köklerinin Sözcük Bankasında Bulunma Yüzdeleri

Tablo 5: Üretilen Sözcüklerin Kullanım Yüzdesi

Harf Adedi	Üretilenlerin Kullanım Yüzdesi
2	%42,79
3	%33,47
4	%8,63
5	%1,73
6	%0,20
7	%0,02
	ORTALAMA : %0,075

V.SONUÇ

Gerçeklenen çalışma sonucunda dil bilimciler için olası Türkçe sözcük kökleri üreten bir yazılım geliştirilmiştir. Bu şekilde yabancı sözcüklerin dilimize yerleşmeden karşılıklarının bulunması daha kolay olacaktır.

Çalışmamızın iyileştirilmesi adına atılması gereken adımlardan bir tanesi, karşılaştırmalarda kullanılan sözcük bankasının daha fazla sözcük içerecek şekilde güncellenmesidir. Dilbilimciler tarafından bir sözcük kökleri sözlüğünün oluşturulması, bu konuda ileride yapılacak çalışmalar için çok yararlı olacaktır.

Yabancı sözcüklere Türkçe karşılık bulma konusunda bir sonraki çalışma olarak, üretilen sentetik sözcük köklerine, uygun yapım eklerini kurallı bir düzende ekleyerek yeni sözcükler türetilmesi düşünülmektedir. Bu sayede üretilen sözcük köklerinden bir ön eleme yapılarak seçilen köklerden yeni sözcükler türetilmesi gündemdedir.

VI.KAYNAKLAR

[1] Hengirmen, M., Türkçe Dilbilgisi, Engin Yayıncılık, Ankara, 2002.

[2] Banguoğlu, T., Türkçenin Grameri, Türk Dil Kurumu Yayınları : 528, Ankara, 2000.

[3] Keçeci, H. F., Bir Robot Koluna Kumanda Eden Doğal Dil Anlama Sistemi, İstanbul Teknik Üniversitesi, İstanbul, 1996.

[4] Cebiroğlu, G. Adalı E., Sözlüksüz Köke Ulaşma Yöntemi, TBD 19. Bilişim Kurultayı, İstanbul, 2002.

[5] Oflazer, K., PC-KIMMO Rule Specification For Turkish Morphology, Bilkent University, Ankara, 1994.

[6] Türk Dil Kurumu, İmla Kılavuzu, Türk Dil Kurumu Yayınları, Ankara, 2000.