Uluslararası Sosyal Araştırmalar Dergisi / The Journal of International Social Research
Cilt: 12 Sayı: 62 Yıl: 2019 Volume: 12 Issue: 62 Year: 2019
www.sosyalarastirmalar.com Issn: 1307-9581
http://dx.doi.org/10.17719/jisr.2019.3025

SORANÎ KÜRTÇESİNE KARŞI KURMANCÎ KÜRTÇESİ: DENEYSEL BİR KARŞILAŞTIRMA*

SORANI KURDISH VERSUS KURMANJI KURDISH: AN EMPIRICAL COMPARISON

Osman ASLANOĞLU **

Öz

Dil biliminde ve dil işleme çalışmalarında corpus kelimesi ile kastedilen, çok sayıdaki metnin düzenli ve yapısal olarak bir arada bulunması durumudur. Corpuslar tek dilli veya çok dilli olabilir. Corpus yöntemiyle dil metinlerinin pek çok açıdan analizini kolayca yapmak mümkün olmaktadır. Çevirisini yaptığımız bu makalede de Kürtçenin Kurmancî ve Soranî lehçelerinde yapılan kıyaslama sonuçları üzerinde durulmaktadır.

Lehçe ve yazımdaki çeşitlilikle birlikte kaynak azlığı, Kürtçenin işleyişindeki iki ana sorundur. Bu çalışmada Kürtçenin iki ana Lehçesi olan Soranî ve Kurmancî için (i)metin corpusu oluşturarak (ii)istatistiksel ve kurala dayalı bakış açılarıyla bu iki lehçe arasındaki bazı imlasal, fonolojik ve morfolojik farklılıkların önemini vurgulayarak bu iki probleme değinmeyi amaçlıyoruz.

Anahtar Kelimeler: Sorani, Kurmanci, Corpus, Alfabe, Kürtçe, Karşılaştırma.

Abstract

In linguistics and language processing studies, with the term corpus that is implied the situation of coexisting of numerous texts as regularly and structurally. Corpuses may be monolingual or multilingual. With the corpus method, it's possible to analyse language texts easily in many ways. In this article, which we translated, the outcomes of making comparison in the Kurmanji and Sorani dialects of Kurdish are emphasized.

Resource scarcity along with diversity–both in dialect and script–are the two primary challenges in Kurdish language processing. In this paper we aim at addressing these two problems by (i) building a text corpus for Sorani and Kurmanji, the two main dialects of Kurdish, and (ii) highlighting some of the orthographic, phonological, and morphological differences between these two dialects from statistical and rule-based perspectives.

Keywords: Sorani, Kurmanji, Corpus, Alphabet, Kurdish, Compare.

1. Giriş

20-30 milyon civarı ana dilini konuşan insanlar olmasına rağmen (Haig and Matras, 2002; Hassanpour et al., 2012; Thackston, 2006b; Thackston, 2006a), Kürtçe internette mevcut olan tek dilbilimsel kaynağın işlenmemiş metin olarak en az kaynaklı diller arasında yer almaktadır(Walther and Sagot, 2010).

Kaynak azlığı sorununun yanı sıra Kürtçenin farklılığı -hem lehçe hem de yazı dizgesinde- Kürtçenin işleyişindeki bir diğer ana zorluktur (Gautier, 1998; Gautier, 1996; Esmaili, 2012). Aslında Kürtçe, bi-standart(*iki lehçeli dil)* olarak düşünülmektedir(Gautier, 1998; Hassanpour et al., 2012): Arap alfabeyle yazılan Soranî lehçesi ve Latin alfabesiyle yazılan Kurmancî lehçesi. Bu iki lehçeyi ayıran özellikler fonolojik, sözcüksel ve morfolojik yapılardır.

Bu yazıda Kürdistan Üniversitesindeki Kürtçenin işleyişindeki bu iki zorluğu ele almayı amaçlayan bir projenin[1] ilk sonuçlarını rapor edeceğiz. Bu yazıda Özellikle:

1. Kürtçe dilinin ilk kısmen kapsamlı ve alenen mevcut metin corpusunun oluşumunu rapor ediyoruz,

2. Soranî Kürtçesi ve Kurmancî Kürtçesi arasındaki imlasal, fonetik ve morfolojik farklılıklardaki bazı kavrayışları sunuyoruz.

Bu yazının geri kalanı aşağıdaki gibi düzenlenmiştir. İkinci bölümde, ilk önce kısa bir şekilde Kürtçeyi ve iki ana lehçesini tanıtıyoruz, sonra kurallara dayalı bir (a.k.a. corpus independent (bağımsız corpus)) bakış açısından iki lehçenin farklılıklarının önemini vurguluyoruz. Daha sonra, 3.bölümde Pewan metin corpusunu sunduktan sonra, 4.bölümde bunu iki lehçe arasındaki istatistiksel bir karşılaştırmayı yapmak için kullanıyoruz. Bu yazı 5.bölümde sonlandırılıyor.

---

* Bu makale "Sorani Kurdish versus Kurmanji Kurdish: An Empirical Comparison" başlığıyla Kyumars Sheykh Esmaili, Shahin Salavati, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 300–305, Sofia, Bulgaria, August 4-9 2013. c 2013 Association for Computational Linguistics'de yayınlanmış olup, yazarlarının izniyle Türkçeye çevrilmiştir.

** Dr. Öğr. Üyesi, Dicle Üniversitesi, Edebiyat Fakültesi, Doğu Dilleri ve Edebiyatları Bölümü, Kürt Dili ve Edebiyatı Anabilim Dalı, aslanogluosman@gmail.com.

[1] https://eng.uok.ac.ir/esmaili/research/klpp/en/main.htm.

2. Kürt Dili ve Lehçeleri

Kürtçe Hint-Avrupa dillerinin Hint-İran ailesine aittir. En yakın bilinen akrabası Farsçadır. Kürtçe, Türkiye, İran, Irak ve Suriye sınırlarını kapsayan geniş coğrafi bir alan olan Kürdistan'da konuşuluyor. Irak'ın iki resmi dilinden biridir ve İran'da bölgesel bir konuma sahiptir.

Kürtçe zengin lehçeli bir dildir, bazen lehçe sürekliliği olarak anılır (Matras and Akin, 2012; Shahsavari, 2010). Bununla birlikte, bu makale de, Kürtçe dilinin iki yakından alakalı ve yaygın olarak konuşulan Soranî ve Kurmancî lehçelerine odaklanacağız. Bununla birlikte bunlar, Kürtçeyi ana dili olarak konuşanların %75'inden fazlasını oluşturmaktadırlar (Walther and Sagot, 2010).

Aşağıda özetlendiği gibi, bu iki lehçe sadece birtakım dilbilimsel yönlerden değil aynı zamanda yazı sistemi yönüyle de farklılık gösterirler.

2.1. Morfolojik Farklılıklar

Önemli morfolojik farklılıklar (MacKenzie, 1961; Haig and Matras, 2002; Samvelian, 2007):

1. Kurmancî her iki cinsiyeti (kadın: erkek) ve isimler ve zamirlerin[2] zıtlık halini(mutlak: dolaylı) tespit etmede daha dikkatlidir. Soranî bu sistemi büyük ölçüde terk etmiştir ve hallerin görevini yüklenmedeki zamirsel sonekleri kullanır.

2. Geçmiş zaman geçişli fiillerde, Kurmancî tam ergatif (hem geçişli hem geçişsiz) dizgeye sahiptir[3], ama dolaylı zamirleri kaybolan Soranî, birleştirilmiş sözcüklere başvurur.

3. Soranî'de edilgen ve ettirgen yapı fiil morfolojisi yoluyla oluşturulur, Kurmancî'de ise aynı zamanda sırasıyla "hatin" (gelmek) ve "dan" (vermek) yardımcı fiilleriyle oluşturulabilirler.

4. Belirli belirleyicisi -aka sadece Soranî lehçesinde görülür.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Arapça'ya dayalı	ا	ب	ج	چ	د	ێ	ف	گ	ژ	ک	ل	م	ن	و	پ	ق	ر	س	ش	ت	وو	ڤ	خ	ز
Latinceye dayalı	A	B	C	Ç	D	Ê	F	G	J	K	L	M	N	O	P	Q	R	S	Ş	T	Û	V	X	Z

(a) Bire bir yerleşim

	25	26	27	28
Arapça'ya dayalı	ئ / ی	و	ی	ه
Latinceye dayalı	I	U / W	Y / Î	E / H

(b) Bire iki yerleşim

	29	30	31	32	33
Arapça'ya dayalı	ڕ	ڵ	ع	غ	ح
Latinceye dayalı	(RR)	-	(E)	(X)	(H)

(c) Bire sıfır yerleşim

Şekil 1: İki standart Kürtçe Alfabe

2.2. Yazı Farklılıkları

Jeopolitik sebeplerden dolayı (Matras and Reershemius, 1991), her iki lehçe de kendi yazı sistemini kullanıyor: Soranî, Arapça kökenli bir alfabe kullanırken, Kurmancî Latin kökenli bir alfabeyle yazılır.

Şekil 1, iki standart alfabeyi ve aralarındaki üç sınıfa kategorize edilmiş eşleşmeyi göstermektedir:

- geniş bir karakter altkümesi kapsayan bire-bir eşleştirmeler(Şekil 1a)
- ikiye-bir eşleştirmeler(Şekil 1b); iki yazı sistemindeki kendine özgü anlam belirsizliğini ifade ederler (Barkhoda et al., 2009). Bu iki alfabe arasında karakter (alfabe) değişikliği yaparken, bağlamsal bilgi doğru eşleşmeyi seçmede ipuçları sağlayabilir.
- sıfıra-karşı eşleştirmeler(Şekil 1c); iki farklı alt kategoriye daha ayrılabilirler: (i) sert L ve sert R karakterleri ({ڵ} ve {ڕ}) sadece Soranî Kürtçesinde kullanılıyor[4] ve Soranîyle Kurmancî arasındaki kendine özgü bazı fonolojik farklılıkları gösteriyor, ve (ii) arda kalan üç karakter Soranî'de Arapça'dan ödünç alınan kelimelerde en çok kullanılıyor (Kurmancî'de diğer karakterlerle yaklaşıktır).

Bu iki yazı sisteminin fonetik olduğuna dikkat edilmelidir (Gautier, 1998); yani, sesli harfler açık bir şekilde gösteriliyor ve kullanımları zorunludur.

3. Pewan Corpusu(taraması)

Metin corpusları Bilişimsel Dilbilimleri ve Doğal Dil İşleme için gereklidir. Corpus (Gautier, 1998)ve sözlük (Walther and Sagot, 2010), oluşturmadaki bir kaç girişime rağmen Kürtçe hala geniş çapta ve güvenilir genel veya ilgili alana özgü bir corpusa sahip değildir.

---

[2] Cinsiyet ayrımının Kurmancî'nin bazı ağızlarında zayıflatıldığını gösteren kanıt olmasına rağmen(Haig and Matras, 2002).

[3] Son araştırma Kurmancî'deki ergativite içten uyarılmış, değişim veya Türkçeyle temas halinde olmasından dolayı zayıflatıldığını gösteriyor(Dixon, 1994; Dorleijn, 1996; Mahalingappa, 2010), belki de tamamen ismin yalın hali sistemine doğru ilerliyordur.

[4] Bununla birlikte Kurmancî'de de ikinci harf olan sert "R" bulunan az sayıda kelime vardır.

Kürdistan üniversitesinde (UoK), Kürt dili için metin corpusu oluşturmak amacıyla TREC (TREC, 2013) 'in genel uygulamasını takip ettik ve haber makaleleri kullandık. Bir dizi seçenek araştırdıktan sonra iki tane online haber ajansı seçtik: (i) Irak Kürdistanı merkezli popüler çok dilli haber ajansı Peyamner (Peyamner, 2013), ve (ii) Soranî (VOA, 2013b) ve Kurmancî (VOA, 2013a) Amerika'nın sesi web siteleri. Ana seçim kriterlerimiz: (i) makale sayısı, (ii) konu çeşitliliği, ve (iii) gezinme kolaylığı.

Her ajans için, makaleleri getirmesi ve metinsel bağlamlarını çıkarması için bir paletli araç geliştirdik. Peyamner durumunda, makalelerin dil etiketi olmadığı için, ilaveten dile özgü karakterlerin ortaya çıkmasına dayalı her sayfanın diline karar veren basit bir sınıflandırıcı uyguladık.

Özellik		**Soranî Corpus**	**Kurmancî Corpus**
Yazı sayısı	VOA'dan	18.420	5.699
	Peyamner'den	96.920	19.873
	toplam	115.340	25.572
Farklı kelime sayısı		501.054	127.272
Toplam kelime sayısı		18.110.723	4.120.027
Toplam karakter sayısı		101.564.650	20.138.939
Ortalama kelime uzunluğu		5.6	4.8

Tablo 1: Pewan Corpus Temel İstatistikleri

Toplamda, 115, 340 Soranî makalesi ve 25, 572 Kurmancî makalesi toplandı[5]. Makaleler 2003 ve 2012 yılları arası tarihlidir ve boyutları 1KB dan 145KB'a (ortalama 2.6KB) arasında değişmektedir. Tablo 1 Pewan - Kürtçe kelime anlamı "ölçü"- olarak adlandırdığımız corpusumuzun önemli özelliklerini özetlemektedir.

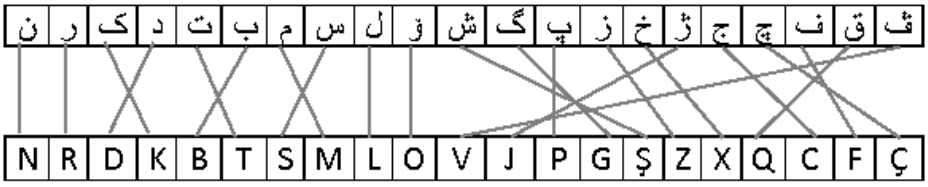
Pewan ve içinde kullanılan benzer yaklaşımı kullanarak (Savoy, 1999), ayrıca Kürtçe gereksiz kelimeler listesi oluşturduk. Bu amaçla, şahsen her bir lehçenin en sık kullanılan 300 kelimesini inceledik ve corpusa özgü önyargıları çıkardık (örneğin, "Irak", "Kürdistan", "Bölgesel", "Hükümet", "Rapor etti" ve benzeri). Son Soranî ve Kurmancî listeleri sırasıyla 157 ve 152 kelime içermektedir ve diğer dillerde olduğu gibi, çoğunlukla edatlardan oluşmaktadır.

Pewan, gereksiz kelime listelerinin yanı sıra (Pewan, 2013)'den elde edilebilir. Bu kaynakların halka açık olmasını sağlamanın Kürt dilinde daha fazla araştırmaya destek olacağını umuyoruz.

4. Deneysel Çalışma

Bu bölümün ilk kısmında, ilk olarak karakter ve kelime sıklıklarını inceliyoruz ve Soranî ve Kurmancî arasında fonolojik ve sözcüksel bağlantılar ve çelişkiler hakkında birkaç öngörü elde etmeye gayret ediyoruz.

İkinci bölümde, iki bilinen dilbilimsel kanunu -Heaps' ve Zibf's-'i inceliyoruz. Bu kanunların çoğu Hint-Avrupa dillerinde (L¨u et al., 2013) incelenmesine rağmen, sonuca etki eden faktörleri dile bağlıdır(Gelbukh and Sidorov, 2001) ve bu yüzden dillerin benzerlik/farklılıklarını ölçmek için bir araç olarak kullanılabilirler. Ayrıca uygulamada bu yasaların veritabanı boyutunun bir fonksiyonu olarak içindekiler kısmının bazı özelliklerini tahmin etmeye olanak sağladığı için sonuca etki eden faktörlerini bilmenin, örneğin, tüm-metin veritabanı tasarımında, önemli olduğu dikkate alınmalıdır.



Şekil 2: Soranî ve Kurmancî karakterlerin Pewan corpustaki yakın frekansları

[5] Kurmancî derlemenin nispeten küçük boyutu daha genel bir akımın parçasıdır. Doğrusu, çok sayıda sözcüye sahip olmasına rağmen, Kurmancî çok daha az kolaylıkla elde edilebilen hazır işlenmemiş metinli çevrimiçi kaynağa sahiptir ve hatta bu kaynaklar tam anlamıyla yazı standartlarını takip etmezler. Bu, kısmen on yıllardır Kurmancî konuşanların çoğunluğunun yaşadığı yer olan Türkiye'deki Kürtçe dilinin kullanımındaki ciddi sınırlamaların bir sonucudur(Hassanpour et al., 2012).

#	English Trans.	Freq.	Sorani Word
1	from	859694	لە
2	and	653876	و
3	with	358609	بە
4	for	270053	بۆ
5	which	241046	کە
6	that	170096	ئەو
7	this	83445	ئەم
8	of	74917	ی
9	together	58963	لەگەڵ
10	made/did	55138	کرد

Kurmanji Word	Freq.	English Trans.	#
û	166401	and	1
ku	112453	which	2
li	107259	from	3
de	82727	-	4
bi	79422	with	5
di	77690	at	6
ji	75064	from	7
jî	57655	too	8
xwe	35579	oneself	9
ya	31972	of	10

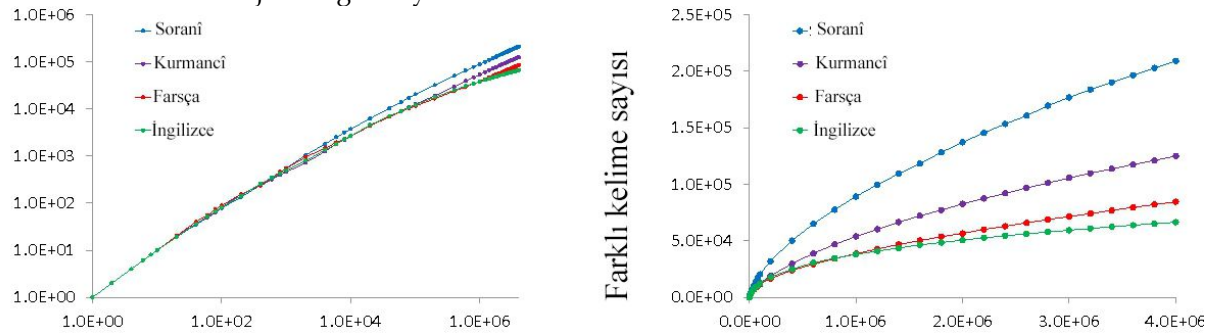
Şekil 3: Pewan'da Kurmancî ve Soranî en çok kullanılan kelimeler.

4.1. Karakter Sıklıkları

Bu deneyde dilin fonolojik bir özelliği olarak karakter sıklığını ölçüyoruz. Şekil 2 Pewan corpusunda her iki lehçenin karakterlerin sık kullanım sırasına ait listeleri (soldan sağa, azalan bir sırada) gösteriyor. Doğru bir karşılaştırma için 1'e 1 eşleşme listesinden üç karakterin yanı sıra 1'e 0 ve 1'e 2 eşleşmesi ile karakterleri dâhil etmedik: A, ˆE ve ˆU. İlk ikisi izafe tümce yapısı[6] belirteci görevinden dolayı çarpık bir frekansa sahiptir. Üçüncüsü ise Soranî alfabesindeki çift karakterle eşleştirildi.

Genel olarak, bu iki listedeki eşdeğer karakterlerin nisbî konumları kıyaslanabilir (Şekil 2). Ancak, Soranî ve Kurmancî arasında kendine has fonolojiksel farklılıkları ayrıca gösteren iki dikkate değer çelişki vardır.

- J karakterinin kullanımı Kurmancî'de daha çok yaygındır (örneğin, ji "den" ve jî "de/da" edatlarındaki gibi),
- aynısı V karakteri için de geçerlidir; bu, ancak, Soranî'nin V yerine W sesbirimini kullanımındaki fonolojiksel eğilimi yüzündendir.



(a) Standart Gösterim (b) Logaritmik olmayan Gösterim

Şekil 4: Soranî ve Kurmancî Kürtçesi, Farsça ve İngilizce için Heaps Kanunu

4.2. Kelime Sıklıkları

Şekil 3 Pewan corpusunda en çok kullanılan Soranî ve Kurmancî kelimelerini göstermektedir. Bu şekil ayrıca başka dilde yazılan eşdeğer kelimeler arasındaki bağlantıları içermektedir ve ayrıca iki lehçe arasındaki yüksek düzeyde bir ilişki gösterir. Sık kullanılan terimler listesinin daha uzun biçiminin derinlemesine bir incelemesi, sadece bu ilişkiyi daha fazla doğrulamaz, ayrıca bazı diğer dikkate değer örnekleri açığa çıkarır:

- Soranî soyuna özgü edat لە("-den") çok kapsamlı bir kullanışa sahiptir: aslında, Şekil 3 de gösterildiği gibi, üç ortak Kurmancî edatının (li, ji ve di) anlamsal eşdeğeridir,
- Soranî'de بوون"olmak" fiilinin yanı sıra bir takım ortak edat (örneğin, یش "-de/da") son-ek olarak kullanılıyor,
- Kurmancî'de en yaygın bazı edatlar sözcük sonuna konan bir ek (çoğunlukla, da, de ve ve) ile birlikte ele alınırlar,
- Kurmancî'nin edilgen/ettirgen yardımcı fiilleri (hatin ve dan) en yaygın kullanılan kelimeler arasında yer alır.

4.3. Heaps Kanunu

Heaps kanunu (Heaps, 1978), belirgin kelimelerin gelişimi hakkındadır (a.k.a kelime boyutu). Özellikle, bir metindeki belirgin kelimelerin sayısı, ölçüsünün katsayısına yaklaşık olarak orantılıdır:

$$\log ni \approx D + h \log i$$

[6] İzafe tümce yapısı birtakım batı İrani dillerin paylaşılan bir özelliğidir(Samvelian, 2006). Aşağı yukarı İngilizce edatı olan "(of)-nın"a benzemektedir ve bir sözcük grubunda edatların, isimlerin ve sıfatların arasına eklenir(Shamsfard, 2011).

Dil	log *ni*	*h*
Soranî	1.91 + 0.78 log *i*	0.78
Kurmancî	2.15 + 0.74 log *i*	0.74
Farsça	2.66 + 0.70 log *i*	0.70
İngilizce	2.68 + 0.69 log *i*	0.69

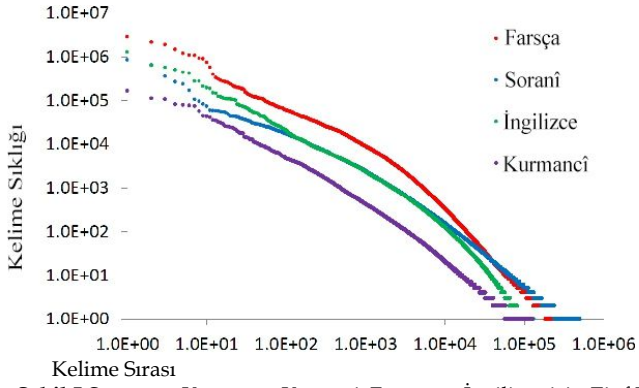
Tablo 2: Heaps'in doğrusal gerilemesi

Burada *ni* sayısı, *i* kelime sayısını kullanmadan önce ortaya çıkan farklı kelimelerin sayısıdır, *h* gösterge katsayısıdır ve *D* ise sabit bir sayıdır. Logaritmik bir ölçüde, yaklaşık 45 açılı düz bir çizgidir (Gelbukh and Sidorov, 2001).

Farsça ve İngilizce dillerinin yanı sıra her iki Kürtçe lehçesinin belirgin kelimelerinin gelişim oranını ölçmek için bir deney yaptık. Bu deneyde, Farsça corpus standart Hamshahri koleksiyonundan hazırlanmıştır (AleAhmad et al., 2009) ve İngilizce corpus Guardian gazetesinin editöryel makalelerinden oluşmaktadır[7](Guardian, 2013).

Şekil 4'teki eğrilerin ve Şekil 2'deki doğrusal bağlanım katsayısının gösterdiği gibi, hem Soranî hem de Kurmancî Kürtçesindeki belirgin kelimelerin gelişim oranı Farsça ve İngilizce'den daha fazladır. Sonuç Kürtçe dilinin morfolojik karmaşıklığını göstermektedir(Samvelian, 2007;Walther, 2011). Bu karmaşıklığın arkasında yatan etkin faktörlerden biri, son-ek'lerin yaygın kullanımıdır, en dikkat çekeni olarak: (i) İzafe tümce yapı belirteci, (ii) çoğul isim belirteci, ve (iii) belgisiz belirteç.

Bu deneydeki bir diğer önemli gözlem Soranî'nin, Kurmancî'ye kıyasla daha yüksek bir gelişim oranına sahip olmasıdır (h = 0.78 vs. h = 0.74).



Şekil 5: Soranî ve Kurmancî Kürtçesi, Farsça ve İngilizce için Zipf Kanunu

Dil	log *f r*	*z*
Soranî	7.69 − 1.33 log r	1.33
Kurmancî	6.48 − 1.31 log r	1.31
Farsça	9.57 − 1.51 log r	1.51
İngilizce	9.37 − 1.85 log r	1.85

Tablo 3: Zipf''in doğrusal gerilemesi

Bu farklılıkların iki başlıca kaynağı: (i) daha öncede bahsedildiği gibi iki lehçenin arasındaki kendine özgü farklılıklar (özellikle, Soranî'nin belirli belirteçlerinin özel kullanımı), (ii)son-ek olarak edatları ve yardımcı fiilleri kullanmada Soranî'deki genel eğilim.

4.4. Zipf Kanunu

Zipf Kanunu (Zipf, 1949), herhangi yeterince büyük bir metinde, kelimelerin sıklık sırası, ilgi sıklığıyla ters orantılıdır:

$$\log fr \approx C - z \log r \quad (2)$$

fr kelime sıklığının üst kat sayı, *r*, *z*'nin üssel katsayısı ve *C* sabit bir katsayıdır. Logaritmik bir ölçüde, yaklaşık 45 açılı düz bir çizgidir(Gelbukh and Sidorov, 2001).

[7] Ana dilini konuşanlar tarafından yazıldıkları için, 2006 ve 2013 yılları arasında geniş bir konu görüntüsü kapsamaktadır ve açık HTML kaynaklarına sahiptir.

Şekil 5'te eğrileri çizilen deneyimizin ve Tablo 3'teki doğrusal bağlam katsayılarının sonuçları gösteriyor ki: (i) Soranî'de en yaygın olan kelimelerin dağılımı benzersiz bir şekilde farklıdır; ilk olarak en yaygın 10 kelimenin keskin bir düşüşünü ve daha sonra 10 ile 100 arasında sıralanan kelimelerin daha yavaş düşüşünü göstermektedir ve (ii) eğrilerin kalan kısımlarında hem Kurmancî hem de Soranî'de benzer şekilde hareket eder; bu aynı zamanda *z* katsayılarında yansıtılmıştır (1.33 and 1.31).

5. Sonuçlar ve Gelecek İş

Bu yazıda Kürtçe dili işleme sürecindeki iki ana zorluğu -isim vermek gerekirse, kaynak kıtlığı ve çeşitliliği- ele alma doğrultusundaki ilk adımları attık. Kürtçenin iki temel lehçesi olan Soranî ve Kurmancî için yapılan bir metin corpusu olan Pewan'ı sunduk. Ayrıca bu iki lehçe arasındaki bir takım farklılıkları ve yazı sisteminin önemini vurguladık.

Analizimizin ana buluşları aşağıdakiler gibi özetlenebilir: (i) Soranî ve Kurmancî arasında fonolojik farklılıklar vardır; bazı sesbirimleri Kurmancî lehçesinde yokken, diğer bazıları Soranî'de daha az yaygındır, (ii) kelime gelişim oranı bakımından oldukça farklılık göstermektedirler, (iii) Soranî kendine özgü bir frekans dağılımına sahiptir w.r.t. son derece yaygın kelimeleridir. Soranî'deki bazı tutarsızlıklar, kendine özgü bir edat olan ﻯ nin varlığı, ve kendi yazı sistemi ve stilinde edatları sonek olarak kullanmadaki genel eğiliminden dolayıdır.

Kürdistan Üniversitesindeki projemiz devam etmekte olan bir çalışmadır. Son zamanlarda, Pewan corpusunu Kürtçe bilgi elde etme sistemini değerlendirmede bir test derlemesi oluşturmada kullandık (Esmaili et al., 2013). Gelecekte, ilk olarak hem Soranî hem de Kurmancî için algoritmaların kökenine indirgemeyi ve daha sonra bu algoritmaları iki lehçenin arasındaki sözcüksel farklılıkları incelemek için kullanmayı planlıyoruz. Gelecekte çalışılacak diğer bir diğer alan ise Soranî ve Kurmancî arasında bir harf (alfabe) çevirisi/çeviri motoru yapmak.

Teşekkür:
Anonim okurlarımıza yazının kalitesini geliştirmede yardımcı olan anlayışlı yorumlarından dolayı şükranlarımızı sunarız.

KAYNAKÇA

Abolfazl, AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2009. Hamshahri: A standard Persian Text Collection. *Knowledge-Based Systems*, 22(5):382–387.

Alexander, Gelbukh and Grigori Sidorov (2001). Zipf and Heaps Laws' Coefficients Depend on Language. In *Computational Linguistics and Intelligent Text Processing*, pages 332–335. Springer.

Amir, Hassanpour, Jaffer Sheyholislami, and Tove Skutnabb-Kangas (2012). Introduction. Kurdish: Linguicide, Resistance and Hope. *International Journal of the Sociology of Language*, 2012(217):118.

David, N. MacKenzie (1961). *Kurdish Dialect Studies*. Oxford University Press.

Faramarz, Shahsavari (2010). Laki and Kurdish. *Iran and the Caucasus*, 14(1):79–82.

George, Kingsley Zipf (1949). *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley.

Geraldine Walther and Benoˆıt Sagot (2010). Developing a Large-scale Lexicon for a Less-Resourced Language. In *SaLTMiL's Workshop on Lessresourced Languages (LREC)*.

Geraldine, Walther (2011). Fitting into Morphologi- ´cal Structure: Accounting for Soranî Kurdish Endoclitics. In Stefan Muller, editor, ¨ *The Proceedings of the Eighth Mediterranean Morphology Meeting (MMM8)*, pages 299–322, Cagliari, Italy.

Gérard, Gautier (1996). A Lexicographic Environment ´for Kurdish Language using 4th Dimension. In *Proceedings of ICEMCO*.

Gérard, Gautier (1998). Building a Kurdish Language ´Corpus: An Overview of the Technical Problems.In *Proceedings of ICEMCO*.

Goeffrey, Haig and Yaron Matras (2002). Kurdish Linguistics: A Brief Overview. *Sprachtypologie und Universalienforschung / Language Typology and Universals*, 55(1).

Guardian (2013). The Guardian. www.guardian.co.uk/.

Harold, Stanley Heaps (1978). *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc. Orlando, FL, USA.

Jacques, Savoy (1999). A Stemming Procedure and Stopword List for General French Corpora. *JASIS*, 50(10):944–952.

Kyumars, Sheykh Esmaili (2012). Challenges in Kurdish Text Processing. *CoRR*, abs/1212.0074.

Kyumars, Sheykh Esmaili, Shahin Salavati, Somayeh Yosefi, Donya Eliassi, Purya Aliabadi, Shownm Hakimi, and Asrin Mohammadi (2013). Building a Test Collection for Soranî Kurdish. In *(to appear) roceedings of the 10th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA '13)*.

Laura, Mahalingappa (2010). The Acquisition of SplitErgativity in Kurmanji Kurdish. In *The Proceedings of the Workshop on the Acquisition of Ergativity*.

Linyuan, Lu, Zi-Ke Zhang, and Tao Zhou (2013). De- ¨viation of Zipf's and Heaps' Laws in Human Languages with Limited Dictionary Sizes. *Scientific reports*, 3.

Margreet ,Dorleijn (1996). The Decay of Ergativity in Kurdish.

Mehrnoush, Shamsfard (2011). Challenges and Open Problems in Persian Text Processing. In *Proceedings of LTC'11*.

Pewan (2013). Pewan's Download Link. https://dl.dropbox.com/u/10883132/Pewan.zip.

Peyamner (2013). Peyamner News Agency. http://www.peyamner.com/.

Pollet, Samvelian (2006). When Morphology Does Better Than Syntax: The Ezafe Construction in Persian. *Ms., Universite de Paris ´*.

Pollet, Samvelian (2007). A Lexical Account of Soranî Kurdish Prepositions. In *The Proceedings of the 14th International Conference on Head-Driven Phrase Structure Grammar*, pages 235–249, Stanford. CSLI Publications.

Robert, MW Dixon (1994). *Ergativity*. Cambridge University Press.

TREC (2013). Text REtrieval Conference. http://trec.nist.gov/.

VOA (2013a). Voice of America - Kurdish (Kurmanji). http://www.dengeamerika.com/.

VOA (2013b). Voice of America - Kurdish (Soranî).http://www.dengiamerika.com/.

Wafa Barkhoda, Bahram ZahirAzami, Anvar Bahrampour, and Om-Kolsoom Shahryari. 2009. A Comparison between Allophone, Syllable, and Diphone based TTS Systems for Kurdish Language. In *Signal Processing and Information Technology (ISSPIT), 2009 IEEE International Symposium on*, pages 557–562.

Wheeler, M. Thackston (2006a). *Kurmanji Kurdish: A Reference Grammar with Selected Readings*. Harvard University.

Wheeler, M. Thackston (2006b). *Soranî Kurdish: A Reference Grammar with Selected Readings*. Harvard University.

Yaron, Matras and Gertrud Reershemius (1991). Standardization Beyond the State: the Cases of Yiddish, Kurdish and Romani. *Von Gleich and Wolff*, 1991:103–123.

Yaron, Matras and Salih Akin (2012). A Survey of the Kurdish Dialect Continuum. In *Proceedings of the 2nd International Conference on Kurdish Studies*.