ULUSLARARASI
DİSİPLİNLER ARASI DİL ARAŞTIRMALARI DERGİSİ
[DADA]

www.dadaakademi.com

Araştırma Makalesi

KUZEYDOĞU TÜRK LEHÇELERİNİN İSTATİSTİKSEL BİR KARŞILAŞTIRMASI

Ezgi ASLAN*

Özkan ASLAN**

Makale Bilgisi

Geliş Tarihi 15.12.2021
Kabul Tarihi 18.12.2021
Yayın Tarihi 22.12.2021

DOI: 10.48147/dada.56

Yazar Bilgisi

** Dr.*

https://orcid.org/0000-0002-0638-280X

ezgicorga@anadolu.edu.tr

Anadolu Üniversitesi, Eskişehir / TÜRKİYE

*** Dr. Öğr. Üyesi*

https://orcid.org/0000-0002-2680-5419

Afyon Kocatepe Üniversitesi, Afyonkarahisar / TÜRKİYE

ÖZ

Türk dillerinin ve lehçelerinin sınıflandırılmasında genel olarak Türklerin yaşadığı bölgelerin coğrafi konumlarına ve ses özelliklerine göre yapılmış sınıflandırmalar ön plana çıkmaktadır. Türklerin tarihindeki göçlerin yoğunluğu ve farklı kültürlerle karşı karşıya gelmiş olması, sınıflandırmayı daha da güçleştirmektedir. Kuzeydoğu grubu lehçeleriyle ilgili sınıflandırmalarda Altay, Hakas ve Tuva Türkçeleri yer alırken kimi sınıflandırmalarda bu gruba Yakutça (Sahaca) da katılmakta (bk. Killi, 2002; Özyetgin, 2006 vd.) kimi sınıflandırmalarda ise Halaççа ve Çuvaşça ile birlikte Türk dilinin uzak lehçelerinden sayılmaktadır (bk. Kirişçioğlu, ve ark. 2018). Yine de Yakutçanın bu grupla olan ilişkisinin altı pek çok çalışmada çizilmiştir. Bu çalışmanın iki amacı vardır. İlki; Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutçanın çeşitli ölçütler açısından karşılaştırılması, ikincisi Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutça tümceleri ile Türkiye Türkçesi tümceleri arasındaki düzenleme uzaklıklarının hesaplanmasıdır. Bu bağlamda Türk dillerinin sınıflandırılması çalışmalarına disiplinler arası bir katkı sağlamak amaçlanmaktadır. Elde edilen bulgulara göre Yakutça sözcükler diğer Türk lehçelerine oranla anlamlı derecede uzundur. Bu durum Yakutçada uzun ünlülerin harf tekrarıyla gösterilmesinden kaynaklanıyor olabilir. İncelenen lehçeler arasında tespit edilen yüksek korelasyonlar, bu lehçelerin coğrafyaya ve diller arası ses denkliklerine dayalı Kuzeydoğu dilleri sınıflandırmasına paralel bir sayısal bulgudur. Yapılan entropi analizi sonucunda en düşük entropi ve en yüksek şaşırma değerleri Yakutça metinler için elde edilmiştir. Bu bulgu, Yakutçanın sınıflandırmalarda hem Kuzeydoğu lehçelerinden biri olarak kabul edilip hem de ayrı tutulmasını destekleyen bir sonuç olarak ortaya çıkmıştır. Ayrıca Levenshtein uzaklığı açısından yapılan inceleme de bu görüşle uyuşmaktadır. Ayrıca çok boyutlu ölçekleme analiziyle ortaya çıkan şekil, diğer bulguların çoğunluğu ile aynı sonucu vermiş, Yakutçanın; Altay Türkçesi, Hakas Türkçesi ve Tuva Türkçesine kıyasla anlamlı derecede farklı olduğu tezini güçlendirmiştir.

Anahtar Kelimeler

Kuzeydoğu Türk lehçeleri, Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi, Yakutça, çağdaş Türk lehçeleri, entropi, çok boyutlu ölçekleme, düzenleme uzaklığı.

Atıf Bilgisi

Aslan Ezgi, Aslan Özkan (2021). "Kuzeydoğu Türk Lehçelerinin İstatistiksel Bir Karşılaştırması". *Uluslararası Disiplinler Arası Dil Araştırmaları (DADA) Dergisi,* Sayı: 2021/3, Aralık, s. 79-94.

INTERNATIONAL
JOURNAL OF INTERDISCIPLINARY LANGUAGE STUDIES
[JILS]

www.dadaakademi.com

Research Article

A STATISTICAL COMPARISON OF NORTHEASTERN TURKISH DIALECTS

Ezgi ASLAN*

Özkan ASLAN**

Article info

Submitted 15.12.2021
Accepted 18.12.2021
Published 22.12.2021

DOI: 10.48147/dada.56

Authors info

** Dr.*

https://orcid.org/0000-0002-0638-280X

ezgicorga@anadolu.edu.tr

Anadolu Üniversitesi, Eskişehir / TÜRKİYE

***Dr.*

https://orcid.org/0000-0002-2680-5419

Afyon Kocatepe University, Afyonkarahisar / TÜRKİYE

ABSTRACT

In the classification of Turkish languages and dialects, classifications made according to the geographical location and sound characteristics of the regions where Turks live in general come to the fore. The density of migrations in the history of Turks and the fact that they have come across different cultures make classification even more difficult. While Altai, Khakas and Tuvan Turkish are included in the classifications of the Northeastern group dialects, Yakut (Sakha) is also included in this group in some classifications (see Killi, 2002; Özyetgin, 2006 et al.), and in some classifications, it is considered as one of the distant dialects of the Turkish language together with Halac and Chuvash (see Kirişçioğlu, et al. 2018). Nevertheless, the relationship of Yakut language with this group has been underlined in many studies. This study has two purposes. First; The comparison of Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut language in terms of various criteria, and the second one is to calculate the edit distances between Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut sentences and Turkey Turkish sentences. In this context, it is aimed to provide an interdisciplinary contribution to the classification of Turkish languages. According to the findings, Yakut words are significantly longer than other Turkish dialects. This may be due to the repetition of long vowels in Yakut. The high correlations detected among the studied dialects are a numerical finding parallel to the Northeastern language classification of these dialects based on geography and interlingual sound equivalence. As a result of the entropy analysis, the lowest entropy and the highest perplexity values were obtained for the Yakut texts. This finding emerged as a result supporting both the acceptance of Yakut as one of the Northeastern dialects and keeping it separate in classifications. In addition, the examination made in terms of Levenshtein distance also agrees with this view. In addition, the figure revealed by the multidimensional scaling analysis gave the same result as the majority of the other findings, and strengthened the thesis that Yakut language was significantly different from Altai Turkish, Khakas Turkish and Tuvan Turkish.

Keywords

Northeastern Turkish dialects, Altai Turkish, Khakas Turkish, Tuvan Turkish, Yakut (Sakha), contemporary Turkish dialects, entropy, multidimensional scaling, edit distance.

Cite this article as

Aslan Ezgi, Aslan Özkan (2021). "A Statistical Comparison of Northeastern Turkish Dialects". *International Journal of Interdisciplinary Language [JILS] Studies*, Number: 2021/3, December, p. 79-94.

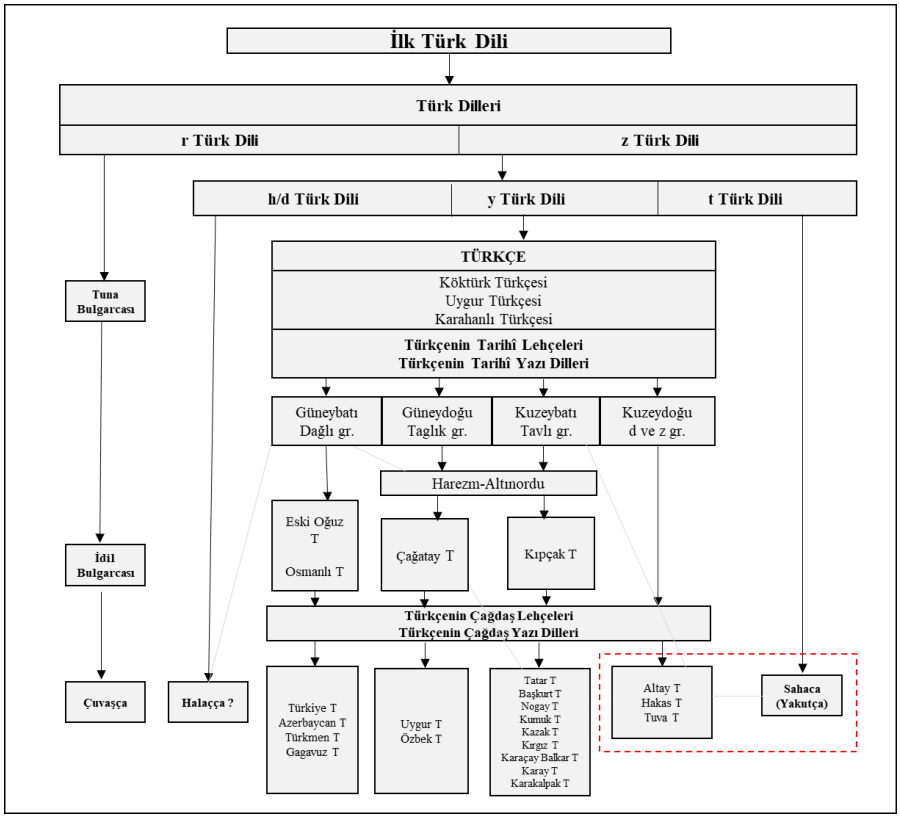
Giriş

Türkçe; dar anlamda Türkiye Cumhuriyeti Devleti'nin resmî dilinin adı, geniş anlamda ise değişik Türk topluluklarının çeşitli lehçeler hâlinde konuştukları ve yazdıkları dilin ortak adıdır (Buran ve ark. 2019, s. 7).

Türk dillerinin ve lehçelerinin sınıflandırılmasında Türkoloji'de bir fikir birliği bulunmamakla birlikte genel olarak Türklerin yaşadıkları bölgelerin coğrafi konumlarına ve ses özelliklerine göre yapılmış sınıflandırmalar ön plana çıkmaktadır. Türk milletinin tarihindeki göçlerin yoğunluğu ve farklı kültürlerle karşı karşıya gelmiş olması, sınıflandırmanın güçlüğünü ortaya koymaktadır. Genel kabul gören coğrafyaya ve ses özelliklerine göre sınıflandırmalardan biri Tablo 1'de verilmiştir.

Türk lehçeleriyle ilgili sınıflandırmalara bakıldığında Kuzeydoğu grubu Türk lehçelerinin d ve z grubuna mensup olduğu görülmektedir. Altay, Hakas ve Tuva Türkçesinin yer aldığı Kuzeydoğu (Sibirya) kolunda ise kimi sınıflandırmalara göre Yakutça (Sahaca) da bulunmaktadır (bk. Killi 2002; Özyetgin, 2006 vd.) ancak Yakutça, birçok bakımdan diğer Kuzeydoğu dillerinden ayrılmaktadır. Dolayısıyla Yakutça, sınıflandırmaların kiminde Çuvaşça ve Halaçça ile birlikte Türk dilinin uzak lehçelerinden biri olarak değerlendirilebilmektedir (bk. Kirişçioğlu ve ark., 2018).

Tablo 1. Türk Dillerinin Sınıflandırması (Buran ve ark., 2019, s. 20)



Entropi ile ilgili çalışmalar, sayısal pek çok alanda uygulandığı gibi sosyal bilimler alanındaki disiplinlerarası çalışmalarda da sıklıkla kullanılmaya başlanmıştır. Entropi optimizasyon metotları; matematik, istatistik, uzay bilimleri, coğrafya, sistem analizi, görüntü işleme, model belirleme, finans, ekonomi, pazarlama gibi pek çok alanda uygulama imkânı bulmuştur. Entropi aynı zamanda metin madenciliği ve doğal dil işlemede de çeşitli analizlerde, semboller/harflerden elde edilen olasılık modellerinin düzensizliğinin hesaplanmasında ve istatistiksel dil modellerinde kullanılmaktadır. Bu

çalışmalardan biri Türkçe lehçelerinin entropilerinin karşılaştırıldığı Saraçlı ve Akın'ın (2014) çalışmasıdır.

1. Kuzeydoğu (Sibirya) Türk Lehçeleri

Farklı Türk boylarının bir arada yaşadığı bir bölge olan Rusya Federasyonu'na bağlı Sibirya'nın güneyinde yaklaşık bir milyon Türk bulunmaktadır ve Kuzeydoğu ya da Sibirya grubu olarak adlandırılan bu boylar, Rusya federasyonuna bağlı Altay, Hakas ve Tuva Cumhuriyetlerinde boylar hâlinde yaşamaktadır (Buran ve ark., 2019, s. 9). Radloff'a göre bütün bu halklarda görünen tek özellik, İslamiyet etkisi altında kalmayan Türklerin küçük bir parçasını oluşturmalarıdır ve bu özellikleri dil ve geleneklerinde de kendini gösterir (Radloff'tan akt. Buran ve ark., 2019, s. 9).

Kendi yazı dilleri ve Rusya federasyonu içinde bağımsız cumhuriyetleri bulunan Saha (Yakut), Tuva, Hakas, Altay Türklerinin dilleri ve ağızları ile ayrı bir cumhuriyete sahip olmayan ve ayrı bir yazı dili bulunmayan Dolgan, Tofa ve Çulım Türklerinin dilleri ile son yıllarda tekrar yazı dili olarak kullanılan Şor lehçesi, Genel Türk dili alanının kuzeydoğusunda bulunduğundan genellikle "Kuzeydoğu grubu Türk lehçeleri" olarak anılagelmişlerdir (Killi, 2002, s. 59). Aynı zamanda Çuvaş Türkçesi dışındaki tüm lehçeler, r/z, l/ş uygunluğuna göre z/ş grubu içinde bulunmaktadır (Killi, 2002, s. 59). Bu çalışmada Altay, Hakas, Tuva Türkçeleri ile Yakutça (Sahaca); kuzeydoğu lehçeleri olarak ele alınacaktır.

1.1. Altay Türkçesi

Altay Türkçesi, Güney-Batı Sibirya'da Rusya Federasyonu'na bağlı Altay Cumhuriyeti ve Altay Kray bölgelerinde yaşayan Altay Türklerinin edebi dilidir (Elcan, 2016, s. 2). Bugün Altaylar Güney Sibirya'da başlıca, %90'ı dağlarla, yarısından fazlası ormanlarla kaplı 92.600 km2'lik bir alanı kapsayan Altay Cumhuriyetinde, Altay Bölgesinde (Kray) ve Kemerovo Bölgesi'nde (Oblast') yaşarlar (Killi Yılmaz, 2011, s. 25).

Altay Türkçesi, zamansal ve mekânsal olarak Türkiye Türkçesinin en uzak akrabalarından biridir (Elcan, 2017, s. 118). Elcan'a göre Altay Türkçesi ile Türkiye Türkçesi arasında görülen ses ve şekil bilgisi farklılıkları iki temel nedene dayanır: Her iki Türk dili tarihsel olarak farklı dil içi gelişim süreçlerinden geçmiş; bu da Eski Türkçe kaynaklı bazı özelliklerin bir lehçede korunurken diğerinde kaybolmasıyla sonuçlanmıştır. Farklılıkların diğer sebebi ise Altay Türkçesi ile Türkiye Türkçesinin tarih boyunca diğer dillerle kurduğu ilişkilerden kaynaklanmaktadır (2017, s. 118).

"Altay" adı Altay-Kiji, Teleüt ve Telengitlerden oluşan güney Altay boylarıyla Tuba, Çalkandu ve Kumandı gibi kuzey Altay gruplarını kapsayacak biçimde bu bölgede yaşayan Türk boylarının/alt etnik gruplarının hepsi için kullanılabildiği gibi son zamanlarda Altayların en kalabalık nüfuslu boyu olan Altay-Kijileri ifade etmek üzere de kullanılmaktadır (Killi Yılmaz, 2011, s. 25). Altay-Kiji ağzına dayanan Altay Türkçesi, Rusça ile birlikte günümüzde de Altay Cumhuriyetinin devlet dilidir.

Altaycanın güney ağızları Oyrot, Telengit, Teleüt ve kuzey ağızları Tuba, Kumandı, Çalkandı'dır. Altayca, Sovyet döneminde 1922'de yazı dili olmuştur. Altayca 1948'e kadar Oyrotça olarak adlandırılmıştır ve günümüzde 52000 (Oyrotça, Teleütçe) kişi tarafından konuşulmaktadır (Özyetgin, 2006, s. 24-25).

1.2. Hakas Türkçesi

Hakas Türkçesi ve ağızları, Rusya Federasyonu'na bağlı Hakas Özerk Cumhuriyetinde (başkent Abakan), Yenisey, Abakan ve Çulım ırmaklarının orta mecrasında yaşayan halklarca konuşulur. Hakas Türkçesinin bir ağzı olan Şor Türkçesi Kemerova eyaletinde konuşulur. Özyetgin'e göre Hakasça konuşan Türklerin asıl çoğunluğunu Sagay-Beltir ve Kaç-Koybal-Kızıl ve Şor grupları oluşturur. Hakas yazı dili bu konuşulan bölgelerde 1944 yılına kadar bir yazı dili olan Şorcanın yerini almıştır (2006, s. 24).

Eski Türkçeye ait birçok özelliği koruyan Hakas Türkçesini ana dili olarak kabul edenlerin sayısı zamanla azalmaya başlamıştır. Bunun nedeni Rusçanın giderek yaygın ve etkin bir şekilde kullanılmasıdır. 1926'da Hakas nüfusunun %96'sı Hakas Türkçesini ana dili olarak kabul ederken 1989'da bu sayı nüfusun %76'sını oluşturmuştur (Buran ve ark., 2019, s. 122).

Yakın geçmişte kaybolan ağızlarıyla birlikte Hakas Türkçesi 6 ağızdan oluşmaktadır: Sagay (Hks. Sağay), Kaç (Haas), Şor (Sor), Beltir (Piltir), Koybal (Hoybal), Kızıl (Hızıl). Bugün ise bu ağızların yalnızca dördü yaşamaktadır (Killi Yılmaz, 2006, s. 83, 85).

Hakas Türkçesi, Rusya Federasyonu'nda Hakas Otonom Cumhuriyetinde 10.000 Şorca olmak üzere 58.000 kişi tarafından konuşulmaktadır.

1.3. Tuva Türkçesi

Tuva Türkçesi, Rusya Federasyonu'na bağlı Tuva Özerk Cumhuriyetinde konuşulan bir Türk lehçesidir. Özkan'a göre Tuvalar arasında Moğolca konuşanların oranı %4,5, Rusça bilenlerin oranı %58,4 Ruslar arasında Tuva Türkçesini bilenlerin oranı ise %0,6'dır. 1989 nüfus sayımında Tuvaların %99,2'si ana dili olarak Tuva Türkçesini bildirmiştir. Bu rakam, Rusya idaresinde bulunan millet ve topluluklar arasında tespit edilen en yüksek ana dilini sahiplenme oranlarından biridir (Özkan'dan akt. Buran ve ark., 2019, s. 200).

Tuva ağızlarıyla ilgili sınıflandırmalardan bugüne kadar geçerli sayılan tasnif Ş.Ç. Sat tarafından yapılmıştır. Sat, Tuva ağızlarını 5 gruba ayırmış, bunların ayırt edici fonetik, morfolojik ve leksik farklılıkları üzerinde durmuştur. Sınıflandırmada coğrafî adlandırmalar kullanılmıştır. Sat'a göre Tuva ağızları: Töp Diyalekt (Merkez Ağzı), Barın Diyalekt (Batı Ağzı), Murnuu-Çöön Diyalekt (Güneydoğu Ağzı), Soñgu-Çöön Diyalekt (Kuzeydoğu Ağzı), Kaa-Hem Ayalgazı ve Tere-Höl Ayalgazı. (Sat'tan akt. Koçoğlu Gündoğdu, 2012, s. 15).

1.4. Yakut (Saha) Türkçesi

Yakut (Saha) Türkçesi, bünyesinde bulundurduğu Moğolca ve Tunguzca unsurlardan dolayı Türk dili ailesi içinde özel bir yer teşkil etmektedir (Kirişçioğlu ve ark., 2018, s. 115). Pek çok sınıflandırmada da Yakut Türkçesi, Türkçenin uzak lehçelerinden biri kabul edilmektedir ancak coğrafi bakımdan bazı sınıflandırmalarda kuzeydoğu dillerine dahil edilmektedir.

Özyetgin'e göre genel Türk dilinin gelişim sürecinden oldukça erken ayrıldığı, coğrafî bakımdan genel Türk dünyasından ayrı düştüğü için bugünkü Türk lehçelerine oldukça uzaktır. Moğol ve Fin-Ugor dilleriyle olan münasebetinden dolayı dilde yabancılaşma oranı da oldukça yüksektir bununla birlikte Yakutça, Türk dilleri içinde en çok Tuvacaya yakındır ancak bununla birlikte iki lehçe arasındaki anlaşabilirlik ölçüsü fazla değildir (2006, s. 24.).

Yakut Türkçesi, Yakutların nüfusu ve siyasi statüleri göz önüne alındığında oldukça iyi durumdadır. Eğitim dili olmamasına rağmen Yakutça bilmeyen Yakut neredeyse yoktur. Bugün Yakut Türkçesi ile kaleme alınmış onlarca gazete ve dergi ile Yakut Türkçesiyle yayın yapan TV ve radyo programları bulunmaktadır (Killi, 2006, s. 64). Dil hususunda oldukça hassas olan Yakutlar, baskın dil Rusça olmasına rağmen kendi aralarında asla Rusçayı kullanmamaktadırlar (Çolak, 2019, s. 5).

Ubryatova'ya göre Hakas Türkçesi, birbiriyle ilişkisi bulunan farklı kabilelerin dillerinin birleşmesi sonucunda oluşmuştur. Yani farklı diller birleşip kaynaşarak genel bir dilin ağızlarını oluşturmuşlardır. Saha Türkçesinde ise tam tersi bir süreç söz konusudur. Saha ağızlarının oluşması, bütün bir yapı gösteren Saha Türkçesinin farklı dilli topluluklara yayılması ile gerçekleşmiştir (Ubryatova'dan akt. Killi Yılmaz, 2006, s. 77).

2. Çalışmanın Amacı ve Yöntemi

Bu çalışmanın iki amacı vardır. İlki; Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutçanın çeşitli ölçütler açısından karşılaştırılması, ikincisi Altay Türkçesi, Hakas Türkçesi, Tuva

Türkçesi ve Yakutça tümceleri ile Türkiye Türkçesi tümceleri arasındaki düzenleme uzaklıklarının hesaplanmasıdır. Bu bağlamda Türk dillerinin sınıflandırılması çalışmalarına disiplinlerarası bir katkı sağlamak amaçlanmaktadır.

Metin madenciliği ve doğal dil işleme süreçlerinde başlangıçta metin verisinin doğasını ortaya çıkarmak amacıyla çeşitli betimleyici istatistikler hesaplanmakta ve metinler üzerinde önişlemleme uygulanmaktadır. Bir metin verisinden veya derlemden çıkarılabilecek en temel veriler; tümce sayısı, sözcük sayısı, tip sayısı ve sembol/harf sayısıdır. Bunların bazıları üzerinde oranlama ve normalleştirme işlemleri gerçekleştirilerek karşılaştırılabilir ölçümler elde edilmektedir. Tip (type) bir dağarcıkta (vocabulary) gözlenen ve yalnızca bir adet bulunabilen sembol dizileridir. Tip sayısının sözcük sayısına bölünmesiyle elde edilen oran tip-sayı oranı (type-token ratio) adı verilmekte olup TSO ile gösterilebilir. Bu oran, ilgili metnin ne ölçüde farklı sözcük içerdiğinin bir ölçüsüdür ve oran büyüdükçe metnin sözcük çeşitliliği açısından zenginleştiği yorumu yapılır. Başka bir oran olarak harf sayılarının sözcük sayısına oranını ifade eden harf-sözcük oranından (HSO) söz edilebilir. Bu oran da sözcük başına düşen harf sayısını verdiği için çoğunlukla ilgili metnin geldiği dilin biçimbilimsel yapısıyla veya kelime seçimleriyle ilgili yorumlar yapmaya izin verir. Aşağıda TSO ve HSO oranlarının formülleri verilmiştir:

$$TSO = \frac{Tip\ Sayısı}{Sözcük\ Sayısı} \qquad HSO = \frac{Harf\ Sayısı}{Sözcük\ Sayısı}$$

Metinleri karşılaştırmalı olarak inceleme yöntemlerinden biri, semboller dizisinin taşıdığı düzensizliği hesaplamaktır. Bu amaçla sembollerin olasılıklarından düzensizlik (entropi) hesaplanır. Entropi genel anlamı itibarıyla bir olayın gerçekleşmesi hakkındaki belirsizliğin ölçüsüdür. Bir sistemin entropisi, sistemin alabildiği bütün durumların belirsizliklerinin toplamıdır. Aşağıdaki formül entropinin genel denklemini ifade etmektedir:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ log\ p(x_i)$$

Örneğin *aaba, aaaa, abaa, babaa,* a*baabaa* gibi beş sözcükten oluşan temsili bir derlem olduğu düşünülürse bu derlemi oluşturan tekil sembollerin toplam entropisi şöyle hesaplanır:

- [p(a) . log p(a) + p(b) . log p(b)]

Bu denklemde p(a) a sembolünün olasılığıdır ve derlemdeki a sembollerinin sayısının toplam sembol sayısına bölünmesiyle elde edilir. Benzer şekilde p(b) de hesaplanırsa:

- [18/24 . log 18/24 + 6/24 . log 6/24] = - [3/4 . log 3/4 + 1/4 . log 1/4]

Logaritma işlemi, enformasyon kuramı ve ikilik sistemle ilişkili olarak genellikle 2 tabanına göre yapılır ve böylece bu örnek derlemin entropisi 2,42 olarak elde edilir. Örnek derlem eşit sayıda a ve b sembolünden oluşsaydı entropi 2 olarak hesaplanacaktı. O hâlde derlemdeki a sembolünün lehine olan baskınlık, düzensizliği artıran bir etken olarak yorumlanabilir.

Entropiye dayalı metin inceleme ve karşılaştırma yapma konusundaki temel sorun, metinlerin yeterli çeşitlilik ve zenginlikte olup olmaması meselesiyle ilişkilidir. Bu konuda şu miktarda metin gereklidir gibi bir varsayımda bulunmak güçtür (Bentz vd., 2017).

Entropinin yanı sıra bir olasılık dağılımının örnekleri ne kadar iyi tahmin ettiğini bir ölçüsü olan şaşırma (perplexity) değeri de hesaplanabilir. Şaşırma, genellikle olasılık modellerini karşılaştırmak için kullanılmaktadır. Bu değerin küçük olması, olasılık dağılımının örneği tahmin etmede başarılı olduğunu gösterir. İlgili formül aşağıda verilmiştir:

$$\frac{1}{\prod_{i=1}^{n} p(x_i)}$$

Yukarıda verilen örnek derlem için şaşırma değeri 1,75 olarak hesaplanır; sembol sayılarının eşit olması senaryosunda ise şaşırma değeri 2 olarak elde edilir. Bu sonuca göre örnek derlemdeki a sembollerinin fazlalığı, entropiyi arttırmakta ancak daha az şaşırma üretmektedir.

Paralel derlemi oluşturan tümce çiftlerinin arasındaki metinsel benzerliği ölçebilmek için çeşitli düzenleme uzaklık ölçüleri kullanılabilir. Bunlardan en yaygın olanları Levenshtein uzaklığı ve en uzun ortak dizidir (longest common subsequence). Levenshtein uzaklığı, bir sembol dizisini diğer sembol dizisine dönüştürmek için gereken en az düzenleme sayısıdır. Buradaki düzenleme; ekleme, silme veya değiştirme olabilir. En uzun ortak dizi ise iki metnin de ortak olarak içerdiği en uzun sembol dizisini ifade eder. Bu ölçümler, metin boyutundan etkilendiği için normalleştirilmeleri gerekir. Bu bağlamda Levenshtein ölçüsünün iki metnin en büyük olanının uzunluğuna, en uzun ortak dizisi uzunluğu ölçüsünün ise iki metnin en küçük olanının uzunluğuna bölünerek normal değerler elde edilebilir.

Diller arası farklılıklar, çeşitli ölçü ve oranlarla ifade edilebildiği gibi bütün bu ölçülerin bir özeti niteliğinde olan ve özel bir tekniğe dayalı çok boyutlu ölçekleme (multidimensional scaling) analizi ile görselleştirilebilir. Çok boyutlu ölçekleme analizinin genel amacı; her noktanın nesneler veya bireyleri temsil ettiği ve nokta çiftleri arasındaki mesafelerin mümkün olduğu orijinal farklılıkları yansıttığı, genellikle Öklidyen bir uzaydaki noktaların konfigürasyonunu bulmaktır (Cox ve Cox, 2008).

Çok boyutlu ölçekleme ile temel bileşenler analizi yapılarak nesneleri tanımlayan bileşenlerin sayısı azaltılır ve sonuç olarak ulaşılan temel bileşenlerle veri hem özetlenmiş hem de görselleştirmeye uygun biçimde az sayıda bileşenle temsil edilmiş olur.

Çalışmanın Evreni ve Örneklemi

Çalışmanın evrenini Kuzeydoğu Türk dillerinden Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutça metinler ile bu dillerde oluşturulmuş metinlerin Türkiye Türkçesi çevirileri oluşturmaktadır. Çalışmanın örneklemini ise Kültür Bakanlığı elektronik kitap projesinin "Türkiye Dışındaki Türk Edebiyatları Antolojisi" bölümünde yer alan Altay Türkçesi (24. cilt) ve Hakas Türkçesi (25. cilt) metinleri ile bunların Türkiye Türkçesine çevirileri, Arçın (2009)'da yer alan Tuva Türkçesi destanı *Haan Tögüldür* ile Türkiye Türkçesi çevirisi ve Yakutça (Sahaca) *Kıus Debeliye* destanı oluşturmaktadır. 2454 satır Hakas Türkçesi, 1060 satır Altay Türkçesi, 4969 satır Yakutça ve 1005 satır Tuva Türkçesi metin ve Türkiye Türkçesi çevirisi Excel tablosunda sayısallaştırılmıştır.

Tuva Türkçesi ile Yakutça metinler yalnızca destan türündeyken Hakas Türkçesi metinleri şiir ve hikâye türlerinde, Altay Türkçesi metinleri ise yalnızca şiir türündedir.

4. Çalışmanın Sınırlılıkları

Çalışma Yakutça ile birlikte Kuzeydoğu dilleriyle sınırlandırılmıştır.

5. Bulgular

Çalışmanın temel malzemesi olan paralel derlem, toplam 9.487 tümce çifti içermektedir. Bu veri, üç farklı Türkçe lehçesinde (Altay Türkçesi, Hakas Türkçesi ve Tuva Türkçesi) yazılmış metinlerin yanı sıra Yakutça tümcelerin Türkiye Türkçesi çevirilerinden oluşmaktadır. Söz konusu derlem, çalışma çerçevesinde iki amaçla kullanılmıştır: 1) Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutçanın çeşitli ölçütler açısından karşılaştırılması, 2) Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutça tümceleri ile Türkiye Türkçesi tümceleri arasındaki düzenleme

uzaklıklarının hesaplanması. Birincisinde yalnızca seçilen dillerdeki tümceler (kaynak metin) işlenirken, ikincisinde paralel derlem, dolayısıyla tümce çiftleri kullanılmıştır.

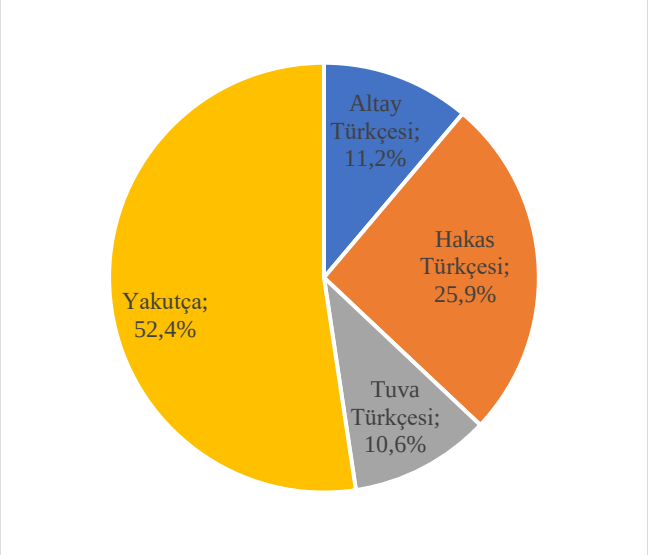
Tablo 2'de paralel derlemden örnek kesitler görülmektedir:

Tablo 2. Paralel derlem kesiti

Lehçe/Dil	S.No	Kaynak Metin	Hedef Metin	Yazar/Derleyen	Tür
Altay T.	61	Edilbegen nemeni	Yapılmayan şeyi	L. V. Kokışev	Şiir
Altay T.	62	Ederge meñdep,	Yapmak için acele edip,	L. V. Kokışev	Şiir
Hakas T.	1	Çir Irağı Kırım'da	Yer Uzağı Kırım'da	K. S. Prokopyeviç	Şiir
Hakas T.	2	İzen, Moskva	Selam, Moskova	K. S. Prokopyeviç	Şiir
Tuva T.	992	Çaa, am bayır eves bayırıp,	Haydi, şimdiye kadar kutlanmayan kutlamayı	S. M. Arçın	Destan
Tuva T.	993	Nayır eves nayırın nayırlan,	Bayram olmayan bayramı	S. M. Arçın	Destan
Yakutça	4918	Körsörbüt baha billibet.	Bir daha karşılaşır mıyız, bilinmez.	S. K. Dyakonov	Destan
Yakutça	4919	Onon oğonńorum bihigini kıtta	Bu yüzden biz yaşlılarla birlikte	S. K. Dyakonov	Destan

Şekil 1'de verileri oluşturan tümcelerin geldiği lehçe/dil oranları verilmiştir:

Şekil 1: Tümce sayısı oranları



Tümce sayısı açısından verilerin yarısından çoğu Yakutça metinlerden gelmektedir. Altay Türkçesi ve Tuva Türkçesi metinleri ise yakın oranlarda olup en az örneği içermektedir. Kaynak metinler üzerinde gerçekleştirilen analizler sonucunda çalışmanın verisiyle ilgili çeşitli istatistikler elde edilmiştir. Söz konusu bulgular Tablo 3'te sunulmaktadır:

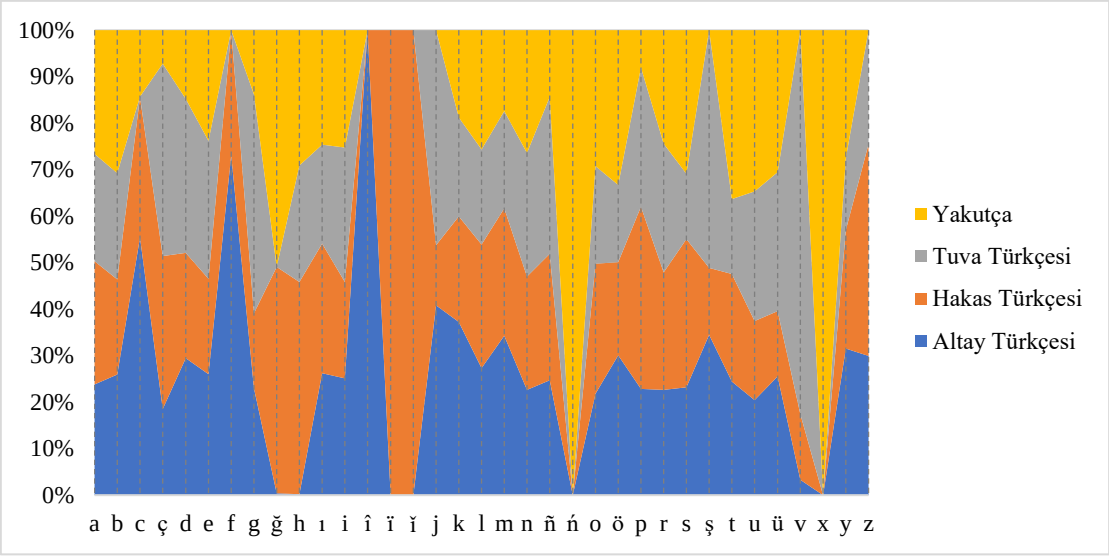
Tablo 3. Betimleyici istatistikler

	Sözcük	**Tip**	**TSO**	**Harf**	**HSO**
Altay Türkçesi	3.392	2.075	0,61	19.896	5,87
Hakas Türkçesi	17.293	7.784	0,45	101.663	5,88
Tuva Türkçesi	4.384	1.631	0,37	24.903	5,68
Yakutça	14.417	5.365	0,37	98.625	6,84
Toplam	39.486			245.087	

Kaynak metinler toplam 39.486 sözcükten oluşmaktadır. Tip sayıları, metinlerden elde edilen dağarcıkların boyutudur. Başka bir ifadeyle tekrar eden sözcüklerin atılmasıyla ortaya çıkan, her sözcüğün yalnızca bir kez gözlendiği listenin uzunluğudur. Örneğin "Ali koşa koşa eve gitti." tümcesinde beş sözcük, dört tip bulunmaktadır. TSO, tip-sözcük oranı ifadesinin kısaltması olup tip sayısının sözcük sayısına bölünmesiyle hesaplanır ve sözcük çeşitliliğinin bir ölçüsüdür. Tablo 3'e göre TSO değeri en yüksek olan 0,61 ile Altay Türkçesi metinleridir. Veriler; sayı, noktalama işaretleri ve boşluklar hariç olmak üzere toplam 245.087 harf içermektedir. HSO, harf-sözcük oranı ifadesinin kısaltması olup harf sayısının sözcük sayısına bölünmesiyle elde edilir. Bu oran 6,84 ile en çok Yakutça için gözlenmiştir.

Kaynak metinlerde 35 farklı harf gözlenmiştir. Büyük harf formları frekans seyrekliğinden kaçınmak için küçük formlara dönüştürülmüştür. Şekil 2'de lehçelere/dillere göre harf olasılıklarının karşılaştırması görülmektedir:

Şekil 2. Karşılaştırmalı harf olasılıkları



Harfler arasında daha çok aksanlı semboller açısından farklılıklar göze çarpmaktadır (bk. Şekil 2.). Bununla birlikte, f ve ğ harfleri Tuva Türkçesinde; j, ş, v ve z harfleri Yakutçada hiç gözlenmemiştir ve x harfiyle yalnızca Yakutça metinlerde karşılaşılmıştır. Dört lehçe/dilin birbiriyle ikili korelasyonları incelendiğinde en düşük Pearson korelasyonu 0,85 ile Tuva Türkçesi ile Yakutça arasında belirlenmiştir. En büyük korelasyon ise 0,94 ile Altay Türkçesi ile Hakas Türkçesi arasındadır.

Betimsel istatistiklere ek olarak, Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutça metinlerinde geçen harflerin tekil entropileri ve şaşırma değerleri hesaplanmıştır. Tablo 4, bu sonuçları göstermektedir:

Tablo 4. Harf entropi ve şaşırma değerleri

	Tek Harf Entropi	Şaşırma
Altay Türkçesi	4,37	80,68
Hakas Türkçesi	4,43	64,91
Tuva Türkçesi	4,36	51,29
Yakutça	4,10	106,63

İlgili analize göre (bk. Tablo 4) en düşük entropi (4,10) ve en yüksek şaşırma değeri (106,63) Yakutça için gözlenmiştir.

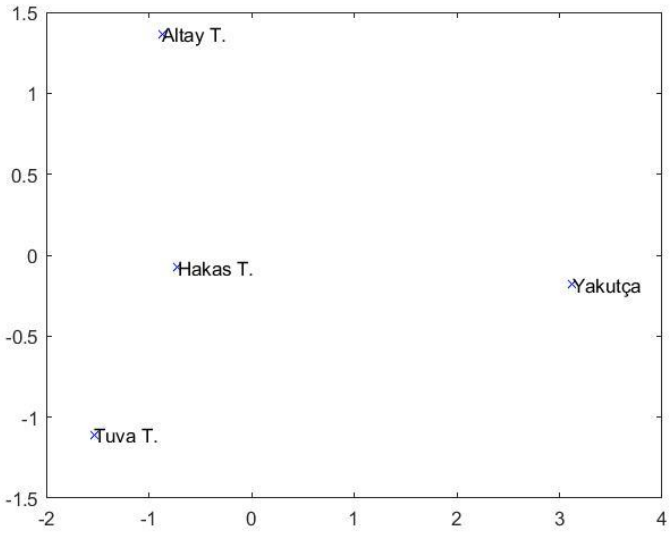
Çalışmanın ikinci amacı olan Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi ve Yakutça tümcelerinin Türkiye Türkçesi çeviri tümceleriyle düzenleme uzaklığı karşılaştırması için paralel derlemdeki tümce çiftleri ele alınmış ve en yaygın düzenleme uzaklığı ölçüleri olan Levenshtein uzaklığı ve en uzun ortak dizi ile değerler hesaplanmıştır. Bu hesaplama sonucunda elde edilen sonuçlar Tablo 5.'te paylaşılmaktadır:

Tablo 5. Düzenleme uzaklığı sonuçları

	Levenshtein uzaklığı	En uzun ortak dizi
Altay Türkçesi	0,65	0,20
Hakas Türkçesi	0,64	0,15
Tuva Türkçesi	0,63	0,22
Yakutça	0,79	0,13

Levenshtein uzaklığı açısından Türkiye Türkçesinden en farklı metinler 0,79 değeri ile Yakutça metinleridir. Üç lehçe metinlerinin ise Türkiye Türkçesine benzerlikleri oldukça yakındır. En uzun ortak dizi ölçüsü açısından da yine en az ortaklık (0,13) Yakutça metinlerle gözlenmiştir. Tablo 3'te yer alan TSO ve HSO ölçüleri ile Tablo 4'teki entropi ve şaşırma ölçülerinden oluşan veri tablosu üzerinde standartlaştırılmış Öklid uzaklığı esasına göre bir uzaklık matrisi elde edilmiş ve bu matris yardımıyla da çok boyutlu ölçekleme analizi uygulanmıştır. Sonuçlar Şekil 3.'te görülmektedir:

Şekil 3. Çok boyutlu ölçekleme analizi



Şekil 3'te verilen grafik, çok boyutlu ölçekleme analizi sonuçlarına göre varyansın yaklaşık %90'ını açıklayan iki temel bileşenin eksenleri olarak almaktadır. Bu bileşenler arasında da en açıklayıcı olan (yaklaşık %75) yatay eksene tekabül etmektedir. Buna göre iki öbekten söz etmek mümkündür: 1) Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi grubu ile Yakutça. İkinci temel bileşene (dikey eksen) göre de daha zayıf bir şekilde üç lehçenin üç ayrı küme oluşturduğu görülmektedir.

6. Yorumlar

Yakutça sözcükler diğer Türk lehçelerine oranla anlamlı derecede uzundur. Bu durumun temel nedeni, Kirişçioğlu ve ark. (2018)'de de belirtildiği gibi "Ana Türkçede olduğu düşünülen ve Türkçenin tarihî dönemlerinde ve çağdaş lehçelerinde görülen aslî uzun ünlülerin büyük bir çoğunluğunu koruyan bir lehçe olması"dır. Örneğin "Kııhırar xaana" (Öfkelenmiş kanı) dizesinde ilk hecelerdeki uzun ünlü gösterimleri, aynı harfin tekrarıyla sağlandığından verinin genelindeki farklılığa yol açmış olabilir.

Bulgularda belirtilen TSO açısından Altay Türkçesinin ayrıklaşmasının nedeni, metin konularının çeşitliliği olabilir nitekim derlemdeki Yakutça ve Tuva Türkçesi metinleri tek bir destandan alınmışken Altay Türkçesi metinleri çeşitli konuları işleyen şiirlerden oluşmaktadır.

İncelenen lehçeler arasında tespit edilen yüksek korelasyonlar (bk. Şekil 2.), bu lehçelerin coğrafyaya ve diller arası ses denkliklerine dayalı Kuzeydoğu dilleri sınıflandırmasına paralel bir sayısal bulgudur. Burada Tuva Türkçesi ve Yakutça arasında ölçülen 0,85 korelasyonunun, Yakutçanın Kuzeydoğu dillerine ait olduğu iddiasını desteklediği söylenebilir.

Tablo 4.'te görüldüğü üzere en düşük entropi ve en yüksek şaşırma değerleri Yakutça metinler için elde edilmiştir. Bu bulgu, Yakutçanın sınıflandırmalarda hem Kuzeydoğu lehçelerinden biri olarak kabul edilip hem de ayrı tutulmasını destekleyen bir sonuç olarak ortaya çıkmıştır. Ayrıca Levenshtein uzaklığı açısından yapılan inceleme de bu görüşle uyuşmaktadır (bk. Tablo 5.).

Çok boyutlu ölçekleme analiziyle ortaya çıkan şekil, önceki bulguların çoğunluğu ile aynı sonucu vermektedir.

Sonuç

Bu çalışmada elde edilen bulgular, Yakutçanın; Altay Türkçesi, Hakas Türkçesi ve Tuva Türkçesine kıyasla harf-sözcük oranı, entropi ve düzenleme uzaklığı ölçütleri açısından anlamlı derecede farklı bir konumda olduğu söylenebilir. Bu sonuçlar, alanyazında yer alan sınıflandırmaları destekleyici niteliktedir (bk. Kirişçioğlu, ve ark. 2018; Buran ve ark., 2019 vd.).

KAYNAKLAR

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. Entropy, 19(6), 275.

Buran, A., Alkaya, E., Özeren, M. (2019). *Çağdaş Türk Yazı Dilleri 4. Kuzeydoğu Grubu*. Akçağ.

Cox, M. A., & Cox, T. F. (2008). Multidimensional scaling. In Handbook of data visualization (pp. 315-347). Springer, Berlin, Heidelberg.

Elcan, A. (2016). *Altay Türkçesi ile Türkiye Türkçesinin karşılaştırmalı ses ve şekil bilgisi* (doktora tezi). Ardahan Üniversitesi, Sosyal Bilimler Enstitüsü.

Elcan, A. (2017). Altay Türkçesinin ayırt edici ses bilgisi özellikleri. *Karadeniz Uluslararası Bilimsel Dergi*, (35), 123-135.

Killi Yılmaz, G. (2011). Altay Türklerinin Dil Durumu. *Modern Türklük Araştırmaları Dergisi,* 8(3), 24-60.

Killi, G. (2002). Hakas Türkçesinin ağızları (doktora tezi), Ankara Üniversitesi, Sosyal Bilimler Enstitüsü.

Kirişçioğlu, F., Gökdağ, B. A., Ersoy, F. ve Doğan, T. (2018). *Türk dilinin uzak lehçeleri*. Akçağ.

Özyetgin, M. (2006). Tarihten Bugüne Türk Dili Alanı. *(Conference) Chinese Academy of Social Science, Sino-Foreign Relationship Department of Institute of History*, Beijing (CHINA) (23 January 2006). www.eurasianhistory.com.

Saraçlı, S. ve Akın, C. (2014). Intertextual comparison of Turkish dialects via entropy approximation. *Electronic Turkish Studies*, *9*(12), 1-7.

EXTENDED SUMMARY

Turkish; in a narrow sense, it is the name of the official language of the State of the Republic of Turkey, and in a broad sense, it is the common name of the language that different Turkish communities speak and write in various dialects (Buran et al. 2019, p. 7).

Although there is no consensus in Turcology in the classification of Turkic languages and dialects, classifications made according to the geographical location and sound characteristics of the regions where Turks live in general come to the fore. The intensity of the migrations in the history of the Turkish nation and the fact that it has faced different cultures reveals the difficulty of classification. One of the classifications according to generally accepted geography and sound characteristics is given in Table 1.

Looking at the classifications of Turkish dialects, it is seen that the Northeast group Turkish dialects belong to the d and z groups. In the Northeastern (Siberian) branch, which includes Altaic, Khakas and Tuvan Turkish, Yakut language (Sakha) is also found according to some classifications (see Killi 2002; Özyetgin, 2006 et al.), but Yakut language differs from other Northeastern languages in many respects. Therefore, Yakut can be considered as one of the distant dialects of the Turkish language along with Chuvash and Halac in some of the classifications (see Kirişçioğlu et al., 2018).

This study has two purposes. First; The comparison of Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut language in terms of various criteria, and the second one is to calculate the edit distances between Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut sentences and Turkey Turkish sentences. In this context, it is aimed to provide an interdisciplinary contribution to the classification of Turkish languages.

The universe of the study consists of Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut texts from Northeastern Turkic languages and Turkey Turkish translations of texts created in these languages. The sample of the study is the Altai Turkish (Volume 24) and Khakas Turkish (Volume 25) texts included in the "Anthology of Turkish Literatures Outside Turkey" of the Ministry of Culture electronic book project and their translations into Turkey Turkish, Tuvan Turkish in Arçın (2009). The epic consists of Haan Tögüldür and Turkey Turkish translation and Yakut (Saha) Kiıs Debeliye epic. 2454 lines of Khakas Turkish, 1060 lines of Altaic Turkish, 4969 lines of Yakut and 1005 lines of Tuvan Turkish text and Turkey Turkish translation have been digitized in the Excel table.

While Tuvan Turkish and Yakut texts are only in epic type, Khakas Turkish texts are in poetry and story types, Altai Turkish texts are only in poetry type. The study is limited to Northeastern languages along with Yakut.

Parallel corpus, which is the main material of the study, contains a total of 9,487 sentence pairs. This data consists of Turkish translations of Yakut sentences as well as texts written in three different Turkish dialects (Altaic Turkish, Khakas Turkish and Tuvan Turkish). The corpus in question was used for two purposes within the framework of the study: 1) Comparison of Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut in terms of various criteria, 2) Calculating the arrangement distances between Altai Turkish, Khakas Turkish, Tuvan Turkish and Yakut sentences and Turkey Turkish sentences. While only the sentences (source text) in the selected languages were processed in the first, parallel corpus, hence sentence pairs, was used in the second.

The source texts consist of 39,486 words in total. Type numbers are the size of the lexicon obtained from the texts. In other words, it is the length of the list in which each word is observed only once, resulting from discarding repetitive words. For example, "Ali ran home running." There are five words and four types in the sentence. TSO is short for type-word ratio and is calculated by dividing the number of types by the number of words and is a measure of lexical diversity. According to Table 3, the Altai Turkish texts have the highest TSO value of 0.61. Data; The number contains 245,087 letters in total, excluding

punctuation and spaces. HSO stands for letter-to-word ratio and is obtained by dividing the number of letters by the number of words. This ratio was mostly observed for Yakut with 6.84.

Yakut words are significantly longer than other Turkish dialects. The main reason for this situation is Kirişçioğlu et al. As stated in (2018), "it is a dialect that preserves the majority of the main long vowels that are thought to be in the main Turkish language and seen in the historical periods and contemporary dialects of Turkish". For example, the long vowel representations in the first syllables in the string "Kıhırar xaana" (Indignant blood) may have caused the difference in the general data, since it is provided by the repetition of the same letter.

In terms of TSO stated in the findings, the reason why Altai Turkish is disjointed may be the diversity of text subjects, as the Yakut and Tuvan Turkish texts in the corpus are taken from a single epic, while the Altai Turkish texts consist of poems dealing with various subjects.

The high correlations detected among the studied dialects (see Figure 2.) is a numerical finding parallel to the Northeastern language classification of these dialects based on geography and interlingual sound equivalence. Here, it can be said that the 0.85 correlation measured between Tuvan Turkish and Yakut supports the claim that Yakut language belongs to the Northeastern languages.

As seen in Table 4., the lowest entropy and highest surprise values were obtained for Yakut texts. This finding emerged as a result supporting both the acceptance of Yakut as one of the Northeastern dialects and keeping it separate in classifications. In addition, the examination made in terms of Levenshtein distance is also compatible with this view (see Table 5.).

The figure revealed by the multidimensional scaling analysis gives the same result as the majority of previous findings.

Keywords

Northeastern Turkish dialects, Altai Turkish, Khakas Turkish, Tuvan Turkish, Yakut (Sakha), contemporary Turkish dialects, entropy, multidimensional scaling, edit distance.

GENİŞ ÖZET

Türkçe; dar anlamda Türkiye Cumhuriyeti Devleti'nin resmî dilinin adı, geniş anlamda ise değişik Türk topluluklarının çeşitli lehçeler hâlinde konuştukları ve yazdıkları dilin ortak adıdır (Buran ve ark. 2019, s. 7).

Türk dillerinin ve lehçelerinin sınıflandırılmasında Türkoloji'de bir fikir birliği bulunmamakla birlikte genel olarak Türklerin yaşadıkları bölgelerin coğrafi konumlarına ve ses özelliklerine göre yapılmış sınıflandırmalar ön plana çıkmaktadır. Türk milletinin tarihindeki göçlerin yoğunluğu ve farklı kültürlerle karşı karşıya gelmiş olması, sınıflandırmanın güçlüğünü ortaya koymaktadır. Genel kabul gören coğrafyaya ve ses özelliklerine göre sınıflandırmalardan biri Tablo 1'de verilmiştir.

Türk lehçeleriyle ilgili sınıflandırmalara bakıldığında Kuzeydoğu grubu Türk lehçelerinin d ve z grubuna mensup olduğu görülmektedir. Altay, Hakas ve Tuva Türkçesinin yer aldığı Kuzeydoğu (Sibirya) kolunda ise kimi sınıflandırmalara göre Yakutça (Sahaca) da bulunmaktadır (bk. Killi 2002; Özyetgin, 2006 vd.) ancak Yakutça, birçok bakımdan diğer Kuzeydoğu dillerinden ayrılmaktadır. Dolayısıyla Yakutça, sınıflandırmaların kiminde Çuvaşça ve Halaççа ile birlikte Türk dilinin uzak lehçelerinden biri olarak değerlendirilebilmektedir (bk. Kirişçioğlu ve ark., 2018).

Entropi ile ilgili çalışmalar, sayısal pek çok alanda uygulandığı gibi sosyal bilimler alanındaki disiplinlerarası çalışmalarda da sıklıkla kullanılmaya başlanmıştır. Entropi optimizasyon metotları; matematik, istatistik, uzay bilimleri, coğrafya, sistem analizi, görüntü işleme, model belirleme, finans, ekonomi, pazarlama gibi pek çok alanda uygulama imkânı bulmuştur. Entropi aynı zamanda metin madenciliği ve doğal dil işlemede de çeşitli analizlerde, semboller/harflerden elde edilen olasılık modellerinin düzensizliğinin hesaplanmasında ve istatistiksel dil modellerinde kullanılmaktadır. Bu çalışmalardan biri Türkçe lehçelerinin entropilerinin karşılaştırıldığı Saraçlı ve Akın'ın (2014) çalışmasıdır.

Yakutça sözcükler diğer Türk lehçelerine oranla anlamlı derecede uzundur. Bu durumun temel nedeni, Kirişçioğlu ve ark. (2018)'de de belirtildiği gibi "Ana Türkçede olduğu düşünülen ve Türkçenin tarihî dönemlerinde ve çağdaş lehçelerinde görülen aslî uzun ünlülerin büyük bir çoğunluğunu koruyan bir lehçe olması"dır. Örneğin "Kıhırar xaana" (Öfkelenmiş kanı) dizesinde ilk hecelerdeki uzun ünlü gösterimleri, aynı harfin tekrarıyla sağlandığından verinin genelindeki farklılığa yol açmış olabilir.

Bulgularda belirtilen TSO açısından Altay Türkçesinin ayrıklaşmasının nedeni, metin konularının çeşitliliği olabilir nitekim derlemdeki Yakutça ve Tuva Türkçesi metinleri tek bir destandan alınmışken Altay Türkçesi metinleri çeşitli konuları işleyen şiirlerden oluşmaktadır.

İncelenen lehçeler arasında tespit edilen yüksek korelasyonlar (bk. Şekil 2.), bu lehçelerin coğrafyaya ve diller arası ses denkliklerine dayalı Kuzeydoğu dilleri sınıflandırmasına paralel bir sayısal bulgudur. Burada Tuva Türkçesi ve Yakutça arasında ölçülen 0,85 korelasyonunun, Yakutçanın Kuzeydoğu dillerine ait olduğu iddiasını desteklediği söylenebilir.

Tablo 4.'te görüldüğü üzere en düşük entropi ve en yüksek şaşırma değerleri Yakutça metinler için elde edilmiştir. Bu bulgu, Yakutçanın sınıflandırmalarda hem Kuzeydoğu lehçelerinden biri olarak kabul edilip hem de ayrı tutulmasını destekleyen bir sonuç olarak ortaya çıkmıştır. Ayrıca Levenshtein uzaklığı açısından yapılan inceleme de bu görüşle uyuşmaktadır (bk. Tablo 5.).

Çok boyutlu ölçekleme analiziyle ortaya çıkan şekil, önceki bulguların çoğunluğu ile aynı sonucu vermektedir.

Anahtar Kelimeler

Kuzeydoğu Türk lehçeleri, Altay Türkçesi, Hakas Türkçesi, Tuva Türkçesi, Yakutça, çağdaş Türk lehçeleri, entropi, çok boyutlu ölçekleme, düzenleme uzaklığı.