21. Metin madenciliği açısından Dede Korkut Kitabı söz varlığının bazı özellikleri

Bekir Tahir TAHİROĞLU[*]

APA: Tahiroğlu, B. T. (2021). Metin madenciliği açısından Dede Korkut Kitabı söz varlığının bazı özellikleri. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, (23), 319-338. DOI: 10.29000/rumelide.948370.

Öz

Dede Korkut Kitabı, Türk dili ve tarihinin önemli yapıtları arasında yer almaktadır. Dede Korkut Kitabı'nın söz varlığı dönemin kültür ve dil özelliklerinin ortaya çıkarılması bakımından da önemli veriler içerir. Tarihsel metinlerin sayısallaştırılmalarıyla hazırlanacak derlemlerin dilbilgisel özelliklere yeni bakış açıları getirmesi yanında dile ait eskiden yeniye söz varlığı değişmelerini de yansıtması, bilgisayar destekli yöntemlerin daha özgül biçimde söz varlığına dair özelliklerin ayrıntılı çıkarımını gerekli kılmaktadır. Dilbilimde söz varlığı incelemeleri çeşitli yöntemlerle yapılabilmekte son yıllarda ise metin ve metin derlemleri temelinde hesaplamalı bir biçimde yeni yöntem ve tekniklerle ele alınabilmektedir. Metin madenciliği temelde yapılandırılmamış bir veri görünümü sunan metinlerden çeşitli örüntülerin elde edilmesi, sınıflandırılması ve makine öğrenmesi tekniklerinin de kullanıldığı yeni gelişen alanlar arasında bulunmaktadır. Çalışmada genel olarak veri madenciliği ve metin madenciliği terimlerinin farklı yönleri ele alınmış ve metin madenciliği bakımından Dede Korkut Kitabı'nın Dresden nüshası esas alınarak nüshadaki bütün sözcüklere ait genel metin istatistikleri, ilk 100 sözcüğün sıklığı, eşdizim ve sözcük ağlarının metin madenciliğinde kullanılan yazılımlar aracılığıyla genel olarak çıkarımları amaçlanmıştır. Sayısallaştırılan metin yazılımların hesaplama modüllerinde yer alan varsayılan istatistik değerleriyle işlenmiş ve elde edilen veriler görsel sonuçlarla da gösterilmiştir. Elde edilen ilk bulgularda 12 hikâyeden oluşan metinde demek, söylemek kavramlarını yansıtan sözcüklerin hem sözcük ağı oluşturmada hem de tekil olarak sıklık listesinde önde gelen sözcükler olduğu görülmüştür. Sonuç olarak, çalışmanın ilk bulgularından hareketle Dede Korkut Kitabı'nın metin madenciliği teknikleriyle daha ayrıntılı söz varlığı ve kavramsal analizinin yapılabileceği ve farklı örüntülerin bulunabileceği düşünülmektedir.

Anahtar kelimeler: Dede Korkut Kitabı, Söz varlığı, metin madenciliği, sıklık, eşdizim, sözcük ağı.

Some features of The Book Of Dede Qorkut vocabulary in terms of text mining

Abstract

The Book of Dede Qorqut is one of the most important works of Turkish language and history. The vocabulary of The Book of Dede Qorkut also reveals important results in terms of revealing the cultural and linguistic characteristics of its period. The fact that the corpora prepared by digitizing the historical texts bring new perspectives to the grammatical features as well as reflect the changes of vocabulary from old to new, necessitates the detailed inference of more specific vocabulary features of computer-aided methods. Analysis of the vocabulary in linguistics can be carry out by

[*] Dr. Öğr. Üyesi, Çukurova Üniversitesi, Fen Edebiyat Fakültesi, Türk Dili ve Edebiyatı Bölümü (Adana, Türkiye) tahirbekir@gmail.com, ORCID ID: 0000-0002-7956-3257 [Araştırma makalesi, Makale kayıt tarihi: 20.04.2021-kabul tarihi: 20.06.2021; DOI: 10.29000/rumelide.948370]

various methods, while in recent years, it can be handled with new methods and techniques in a computational fashion based on text and text collections. In general, different aspects of data mining and text mining terms were discussed in the study and general text statistics of all words in the copy text, frequency of the first 100 words, collocation and lexical networks were generally inferred through software used in text mining, based on the Dresden copy of The Book of Dede Qorkut terms of text mining. Digitized text is processed with the default statistical values contained in the software's calculation modules, and the resulting visual results are presented. In the first findings, it was found that words reflecting the concepts of saying in the text consisting of 12 stories were the leading words in both the word network visualisation and the frequency list. As a result, based on the initial findings of this study, it is believed that a more detailed vocabulary spesific feature and conceptual analysis of The Book of Dede Qorkut can be done using text mining techniques, and thus different patterns can be found.

Keywords: The Book of Dede Qorkut, Vocabulary, text mining, frequency, collocation, word network

Giriş

Dede Korkut Hikayeleri Türk dili ve edebiyatı tarihi ile kültürü açısından en önemli eserler arasında yer almaktadır. Dede Korkut kültürel ve tarihsel olarak Türklerin yaşayış ve düşünüş biçimlerini vermesi ve özellikle destan geleneğinin ortaya konulması açısından önemli bir metindir. Dede Korkut Hikayelerinin yazıya geçirilmesi XV. yüzyıl sonralarına tarihlendirilmektedir (Korkmaz, 1998). Eski Anadolu Türkçesinin özelliklerini taşıyan metnin veri olarak Dresden ve Vatikan nüshaları elimize ulaşmış son olarak da Sahra yazması metninin çeviri yazısı yayımlanarak bilim dünyasına kazandırılmıştır.

Metin madenciliği (*text mining*) metinlerin sayısallaştırılmasından metinlerdeki söz varlığı incelemelerine kadar birçok işlemin otomatik olarak gerçekleştirildiği görece yeni bir alandır. Veri madenciliği (*data mining*) alanıyla birlikte bilgisayar bilimlerinin iki uzmanlık alanı olmakla birlikte son yıllarda dilbilimin alt alanı olan metindilbilim ve söz varlığı araştırmalarında sıklıkla başvurulan alanlar hâline geldiği söylenebilir. Veri madenciliği daha çok sayısal özellikli ya da sayısal biçimde temsil edilen veri setlerinde kullanılan yöntemler bütünü için kullanılırken metin madenciliğinde, başta internet olmak üzere elektronik ortamda yaygınlaşan sözcük içeren yapılandırılmış ya da yarı yapılandırılmış ortamların ayrıntılı incelenmesi, sözcüksel örüntülerin keşfi başta olmak üzere metin birimlerinin yapısal ve anlamsal görünümleriyle ilgilenilmektedir.

Veri madenciliği daha özel olarak da metin madenciliği bilişim sektöründe genellikle kurumsal ihtiyaçların çözümünde kullanılan teknikler bütünü biçiminde görülmektedir. Dilbilim araştırmalarında bu tekniklerin kullanılması, dilsel yapıların da örüntülerden oluştuğu ve istatistiksel davranışlar sergiledikleri düşünüldüğünde, dili ya da yapıtları temsil eden derlemlerden daha önce karşılaşılmamış yapıların, örüntülerin bulunması son derece olası görünmektedir. Bu bağlamda, bu çalışmada metin madenciliği kavramı tanıtılarak uygulamada Dede Korkut Kitabı'nın Dresden nüshasına ait metin verisinden otomatik yöntemlerle elde edilen söz varlığına ait yapıların çeşitli görünümleri verilmiştir. Metinde ilk bakışta bulunması çok zor birliktelikleri, örüntüleri bulmak ve bunları listelemek amaçlanmış, dilbilim araştırmalarında yazılım ve hesaplamalı yöntemlerin kullanılmasının önemi vurgulanmıştır.

1. Veri, metin madenciliği ve söz varlığı

Dede Korkut Kitabı üzerine yapılan oldukça geniş bir yayın çeşitliliği bulunmaktadır. Google Akademik servisinde "dede korkut kitabı"[1] sorgusu için 3570, "dede korkut hikayeleri" sorgusu içinse 2400 sonuç gösterilmektedir. Bu sayılar, eserin ayrıntılı bir biçimde incelenmeye devam edildiğini göstermektedir. Dede Korkut Kitabı'nın hem Dresden hem de Vatikan nüshalarını bir arada yayımlayan Muharrem Ergin, 12 hikâyeden oluşan Dresden nüshasının giriş bölümü ile 12 destan tarzında hikâyeden oluştuğunu belirterek her bir hikâyenin içeriğini vermiştir. Ergin, eseri nitelik açısından da inceleyerek hikâyeleri bir mücadele destanı olarak nitelemiş, bu mücadelelerden ikisinin Oğuz boylarının arasında diğerlerinin de doğa üstü güçlere karşı mücadeleleri içerdiğini belirtmiştir. Asıl metin bölümünden önce hikâyelerin geçtiği coğrafya incelenmiş, Dresden ve Vatikan nüshaları arasındaki farklılıklar bölümünde Vatikan nüshasının eksik bir nüsha olduğu belirterek Dresden nüshasının XVI. yüzyılın ilk yarısında, Vatikan nüshasınınsa XVI. yüzyılın ikinci yarısında istinsah edildiğini vurgulamıştır (Ergin, 1994, s. 1–67).

Veri madenciliği, genel olarak büyük miktarda verinin saklanması ve işlenmesiyle ilgili bir bilim alanı olduğu kadar barındırdığı teknikleri uygulamada daha çok işletmeler kullanmaktadır. U. Tuğba Şimşek Gürsoy, *Veri Madenciliği ve Bilgi Keşfi* adlı kitabında, veri madenciliği kavramından önce veri ambarı (*data warehouse*) terimine açıklık getirmektedir. Bu kavram, özellikle işletmelerin karar alma süreçlerinde veri güdümlü ya da veriye dayalı olarak kullandıkları teknikleri kapsamaktadır. Veri ambarları, çeşitli zamanlarda elde edilmiş parçalı verinin birleştirilmesi ve düzenlenmesiyle ilgilidir. İşletmelerde ya da organizasyonlarda müşterilerden alınan verilerin zamana duyarlı olması, önceye ait verilerden hareketle geleceğe dair tahminlerin yapılmasını, büyük hacimli verinin sayısal özelliklerinden müşterilerin alışveriş davranışlarındaki örüntülerin bulunmasını sağlamaktadır (Gürsoy, 2009, s.3-4). Son zamanlarda veri ambarı teriminin yerini veri madenciliğine bıraktığı söylenebilir. Veri ambarlarında tutulan büyük ölçekli veriler, daha spesifik tekniklerle analiz edilerek ham veriden yeni keşiflerin yapılmasına olanak tanımaktadır. İşte veri madenciliği teknikleri veri tabanı ve veri ambarlarındaki büyük ölçekli veriyi çözümleyip kullanışlı bilgiye dönüştüren süreçlerin adı olarak bilinmektedir.

İnternet verisinin çığ gibi büyüdüğü son yıllarda veri madenciliği alanı önem kazanmış ve uygulama alanlarını genişletmiştir. Silahtaroğlu (2008), veri madenciliği uygulama alanlarını; pazar sepeti analizi, müşteri özelliklerinin çıkarımı, risk yönetimi ve dolandırıcılık tespiti, müşteri değerlendirme, satın alma davranışlarının belirlenmesi olarak başlıklandırmışdır. Alan, bilimsel bilginin derlenmesi ve disipline özgü yöntem bilgisiyle harmanlanarak farklı bakış açılarıyla verinin değerlendirilmesine de olanak tanımaktadır. Örneğin biyoenformatikte DNA analizlerinin yapılmasında ve örüntülerin bulunmasında istatistiksel örüntü tanıma yoğun olarak kullanılabilmektedir.

Verinin elde edilmesinden bilgiye ulaşılıncaya kadar veriyle ilgili izlenen belirli prosedürler bulunmaktadır. Özkan (2008, s. 39) bu prosedürleri verinin temizlenmesi, bütünleştirilmesi, indirgenmesi ve dönüştürülmesi olarak dört aşama olarak sıralamış ve bu aşamaları veri madenciliği tekniklerinin uygulanmasından önceki aşamalar olarak ele almıştır. Kullanılan algoritma ve teknikler yapılacak işin niteliğine ve toplanan verinin özelliklerine göre değişebilmekte birlikte genel olarak sınıflama ve kümeleme analizi gibi iki ana başlık altında değerlendirilmektedir. Altunkaynak (2019, s. 17) kullanılan yöntemleri sınıflandırma, kümeleme, birliktelik ve özellik seçimi biçiminde sıralamıştır. Özellik seçimi (*feature selection*) makine öğrenmesinde de öne çıkan ve yoğun hesaplamalı işlemleri

[1] Sorguların yapıldığı tarih 5.4.2021'dir.

gerektiren yöntemler bütünüdür. Ham veride dolayısıyla büyük veride (*big data*) bulunan ve sınıflandırılmaları sırasında bir bakıma üzerinde hesaplamaların yapılacağı değişkenlerin neler olabileceğinin belirlenmesi özellik seçiminin konusudur. Sınıflandırma, belirlenen özellikler temelinde hedef sınıfa ait olabilecek özellikleri tespit etme işlemidir. Makine öğrenmesinde de önceden belirlenen sınıf etiketine eğitilmiş bir modelden hareketle daha önce karşılaşılmamış birimlerin tahmin edilerek atanması işlemi de bir sınıflandırma algoritmasının uygulanmasıdır. Sınıflandırmada, istatistikte kullanılan doğrusal regresyon, özellikle sayısal değerlerin tahmin edildiği veri setlerinde sıkça başvurulan bir tekniktir. Geçmiş fiyat değerlerinden gelecekteki fiyatların tahmin edilmesi tipik bir doğrusal regresyon problemi olarak verilebilir. Kümeleme analizinde ise, ham verideki birimlerin birbirleriyle olan benzerliklerine göre gruplandırılması söz konusudur. Kümeleme, sınıflandırmaya göre daha genel bir işlem sayılabilir. Birimlerin verideki konumlarına göre uzaklıklarının ölçülmesiyle elde edilen değerlere göre bir araya gelişleri görselleştirilmektedir. Birliktelik analizi ise veride en sık geçen ikili birimlerin (sepet analizinde makarna alanların ketçap alması gibi makarna-ketçap ikilisi) belirlenmesi söz konusudur ve pazarlama sektöründe market raflarının düzenlenmesinde kullanılmaktadır (Altunkaynak, 2019, s. 17-18).

Bilişim çalışmalarında son 10 yılda veri madenciliği ile birlikte öne çıkan alanlardan biri de metin madenciliğidir (*text mining*). Daha önce sözü edilen veri madenciliği araştırma teknik ve yöntemleri metin madenciliğinde de kullanılmakla birlikte, metinlerin yazılı ve sözlü dile dayalı ürünler olması ele alınan teknikleri daha da özelleşmiş duruma getirmiştir. Oğuzlar (2011, p. 6), metin madenciliğini kullanıcılar ile doküman koleksiyonları arasındaki etkileşimli bir süreç olarak tanımlamakta, veri madenciliğinde yapılandırılmış veri biçimleri üzerinde işlem yapılmasına karşın metinlerin yapılandırılmamış biçimlerinin metin madenciliğinin uğraş alanı olduğunu belirtmektedir. Burada yapılandırılmış ve yapılandırılmamış veriden kasıt yapılandırılmış biçimlerdeki özelliklerin (nitelikler) satır ve sütunlardan oluşan bir yapıyla gösterilmesidir. Metin verisinde özelliklerin (sözcükler, özel adlar, sözcük türleri gibi) ayrıca çıkarılıp tablo ya da listeler biçiminde yapılandırılarak gösterilmesi gerekmektedir. Metinlerdeki söz varlığına ait özelliklerin ayrıntılı biçimde ele alınmasından önce derlenen metinlerin bir ön işlemden geçmesi yapılacak çalışmanın sonuçlarını etkilemektedir. Ön işleme (*preprocessing*) sadece metinlerin değil genel olarak çeşitli türden verinin hazırlanmasında kullanılan bir yöntemler bütünüdür. Metinlerdeki sözcüklerin belirlenmesi, lematizasyonlarının yapılması (sözlükbirimlere dönüştürme), noktalama işaretlerinin ayrılması ve optik karakter tanımadan geçen belgelerdeki okuma hatalarının düzeltilmesi birer ön işleme yöntemi olarak görülmektedir, burada da doğal dil işlemenin sözcüklere ve karakter ayırmalara dayanan teknikleri kullanılmaktadır (Oğuzlar, 2011, s. 30–31).

Metin madenciliğinin ortaya çıkışında bilgisayar bilimlerinin rolü daha fazladır denebilir. Metinlerin bilgisayar destekli çözümlenebilmesi için sayısal bir dönüşüm geçirmesi gerekir, böylece dönüşen metin farklı bir biçimde işlenebilir hâle gelerek salt birim sıklıklarından geometrik boyutlandırmaya varan çeşitlikte bir hesaplama ortamına kavuşturulur. Bu anlamda bilgisayar bilimlerinin bakış açısından 1940'lı yıllardan itibaren başlayan doğal dil işleme, bilgisayarlı dilbilim çalışmaları, veri madenciliği ve metin madenciliğinin temel çalışmaları olarak görülmektedir. 1940-2010 ve sonrası tarihler metin madenciliği teriminin yaygınlaşmaya başladığı 1990'lı yıllara kadar bir dizi gelişim süreci olarak; içerik analizi, veritabanlarından bilgi keşfi , gizil anlamsal analiz ve veri madenciliği aşamalarını içermektedir. Günümüzde de büyük veri ve yapay öğrenme terimleri metinlerle ilgili hemen her türlü hesaplamalı çalışmaları kapsar duruma gelmiştir (Anandarajan, Hill ve Nolan, 2019, s. 3–4).

Metin analitiği (*text analytics*) terimi de metin madenciliği yerine kullanılan başka bir terimdir. Metinlerin işlenmesinde analize dayalı süreçlerin daha özgül aşamalarla ifade edilmesi ve her aşamadan sonra bir diğer aşamaya geçilmesi analitik süreçleri ifade etmektedir. Akbıyık (2019, s. 5–6), analitik süreçleri birer fonksiyon olarak nitelemiş; metinlerin toplanması, ön işlemelerin (birimlere ayırma da dahil olmak üzere dilbilgisel ögelere ayırma) yapılması, sözcüklerin seçilmesi ve filtreleme, sözcüklerin vektörlere dönüştürülmesi ve son olarak da konu bulma, kümeleme, sınıflama gibi daha üst düzeyde madencilik işlemlerini sırasıyla süreç akışı olarak belirtmiştir. Belirtilen son aşamaya gelinceye kadar aşamalardan her biri bir diğerine girdi sağlamaktadır.

Metin analizinin bilgisayar destekli ya da hesaplamalı yöntemlerle yapılması metin üzerinde çalışılabilecek her konunun (biçimsel ve anlamsal) algoritmik bir sürece dâhil edilmesi demektir. Otomatik olarak çıkarılabilecek her birim bir keşif sürecini içermektedir. Sözcük türleri gibi önceden belirlenmiş kategorilerin otomatik çıkarımı olabileceği gibi bir kategori adına dayanmadan örüntü oluşturan gizil yapıların çıkarılması da başlı başına ele alınabilecek metinde keşif sürecidir. Bu noktada Anandarajan vd. (2019, s. 2) metin madenciliğinde yapılan işlemin bir tür yüksek kalitede yeni bilgi türetimi olduğunu belirtmektedir. Bu biçimde, metin üzerinde, metinle doğal olarak ilişkili insanların klasik anlamda okuyarak elde edemeyecekleri farklı türden görünümlerin istatistik de dahil hesaplamalı olarak ortaya çıkarılması söz konusudur.

Söz varlığı kavramı bir dile ait sözcüklerin hem biçim hem anlam özelliklerinin hem de sayısal görünümlerinin bir arada tutulduğu bir veri tabanı olarak düşünülebilir. Karaağaç (2013, s. 745), söz varlığı ve söz hazinesi terimini ayrı terimler olarak eserinde göstermiş, söz varlığı maddesinde söz hazinesi terimine göndermede bulunmuştur. Buna göre söz hazinesi, bir dilde yer alan bütün sözleri kapsamaktadır. Gerek kişi gerekse dilsel topluluğun ürettiği söze dair bütün ürünler söz varlığı olarak adlandırılır. İmer, Kocaman ve Özsoy (2011, s. 233), söz varlığını; bir dilin sözlükbirimlerinin tümü olarak tanımlarken, tanımdaki sözlükçe (*lexicon*) terimini de üretici dönüşümsel dilbilgisinin bir ögesi olarak ele almışlardır. Bu kuramdaki bir "depo bileşeni" olan sözlükçe kuram içinde yansıtma ilkesi adı verilen bir yolla görev üstlenir. Sözlükçe terimi İmer, Kocaman ve Özsoy'da (2011, s. 231) lexicon terimine karşılık olarak "sözlükçe" terimi ayrı bir maddebaşı olarak ele alınmıştır. Buradan terimin dilbilimin başka bir kuramında kullanılan ayrı bir kavramı karşıladığı ve genel anlamda sözlük (*dictionary*) yapısından farklı görüldüğü anlaşılmaktadır. Söz varlığı terimini Günay (2018, s. 391) dildeki sözlüksel birimlerin tümü anlamında tanımlarken "sözcükçe" terimini de Fransızca lexique sözcüğüne karşılık olarak kullanmış ve vokabüler yani sözvarlığı ile sözcükçe arasındaki farkı belirtmiştir. Buna göre, sözcükçe, okuma birimi (Fr. *lexie*) terimini oluşturan birimlerin tümünü içermekte ve "sözceleme öznesi"nin kullandığı tüm sözcüklerin tümünü kapsamaktadır. Söz varlığı ise bu tanıma göre daha dar kapsamlı olarak ele alınmakta ve sözceleme öznesinin "kullandığı" tüm sözcüklerden oluşmaktadır. Kısaca sözvarlığı kullanılmış ya da gerçekleştirilmiş olan sözcükçe ise "kullanılmaya hazır" sözcüklerdir denilebilir. Vardar (1998, s. 190), sözcük dağarcığı biçiminde adlandırdığı söz varlığını, birey kullanımı ya da bir derlemde yer alan sözcüklerin tümü olarak belirtmiştir. Bu noktada dilbilim terimlerinin açıklandığı bu üç kaynakta terimin adlandırılışında farklılıklar bulunsa da tanım olarak bir dildeki tüm sözlerin bir arada bulunması ortaktır. Aksan (2018, s. 15) ise söz varlığı denince dildeki sözcükler değil sözcüklerden daha büyük kalıp sözlerin, deyimlerin ve atasözlerinin de anlaşılması gerektiğini söylemiştir. Bu bakımdan Aksan'a göre söz varlığı sadece sözcüklere ve bunların biçimlenişine ait özellikler değil toplumun tüm kavramsal dünyasını ve yaşayış biçimini de içeren bir yapıya sahiptir. Söz varlığı tanımının sınırlarından hareketle milyarlarca sözcüklük bir derlemde hangi ögelerin nasıl çıkarılabileceği, var olan sözlük listelerinde bulunmayan

yeni birimlerin ya da yeni ögelerin nasıl keşfedileceği bir yöntem araştırmaları platformu olarak söz varlığı kavramının dilbilimdeki önemini artırmaktadır.

Bugün söz varlığı araştırmalarında bilgisayar destekli çalışmalar yapılmakla birlikte alan adı olarak metin madenciliği teknik ve yöntemleri yelpazesinin kullanılması da artmaktadır. Bütüncül bir yaklaşım olarak sadece söz varlığındaki öge ve özellik keşiflerinin dilbilim ve metin madenciliği kesişim noktasında bulunduğunu söyleyebiliriz. Özellikle sözcüksel sıklıkların çıkarılmasından sonra bu sıklık istatistiklerinin metinlerden otomatik konu bulmada nasıl yararlanılabileceği, eşdizimli birimlerin metinlerde anahtar kavramları bulmadaki rolleri (metinlerin eşdizimsel dağılım modelleri) söz varlığı araştırmalarının yararlanacağı ve geliştirileceği yeni teknikler olarak görülebilir. Bu çalışmada söz varlığı incelemelerinde kullanılabilecek metin madenciliği yazılımlarının farklı yönleri, *Dede Korkut Kitabı*'na betimleyici biçimde uygulanmaya çalışılmıştır. Metne yönelik Çitgez (2018) tarafından yapılan çalışmada eserin söz varlığı yapı, köken ve anlam bakımından ayrıntılı bir incelemeye tutulmuş, elde edilen sonuçlar sıklık tabloları ve grafiklerle verilmiştir. Bizim çalışmamızın verisi Dresden nüshasına dayanmakta, Çitgez 2018'de ise eser üzerine yapılan diğer çalışmalardan elde edilen sözcüklerin derlendiği belirtilmektedr. Sözcük türlerinin dağılımlarının da verildiği çalışmada anlama dayalı incelemenin ağırlıklı olduğu görülmektedir. Kullanılan yöntem açısından tüm sözcüklerin fişlenerek sınıflandırıldığı belirtilen çalışmadan (Çitgez, 2018, s. 2) farklı olarak metin madenciliği bakış açısıyla ele alınan bizim çalışmamızda tüm sözcüklerin istatistikleri yazılımlar aracılığıyla çıkarılmıştır. Ayrıca eşdizimli birimlerle sözcük ağlarının oluşturulması da hesaplamalı yöntem kullanılarak gerçekleştirilmiştir.

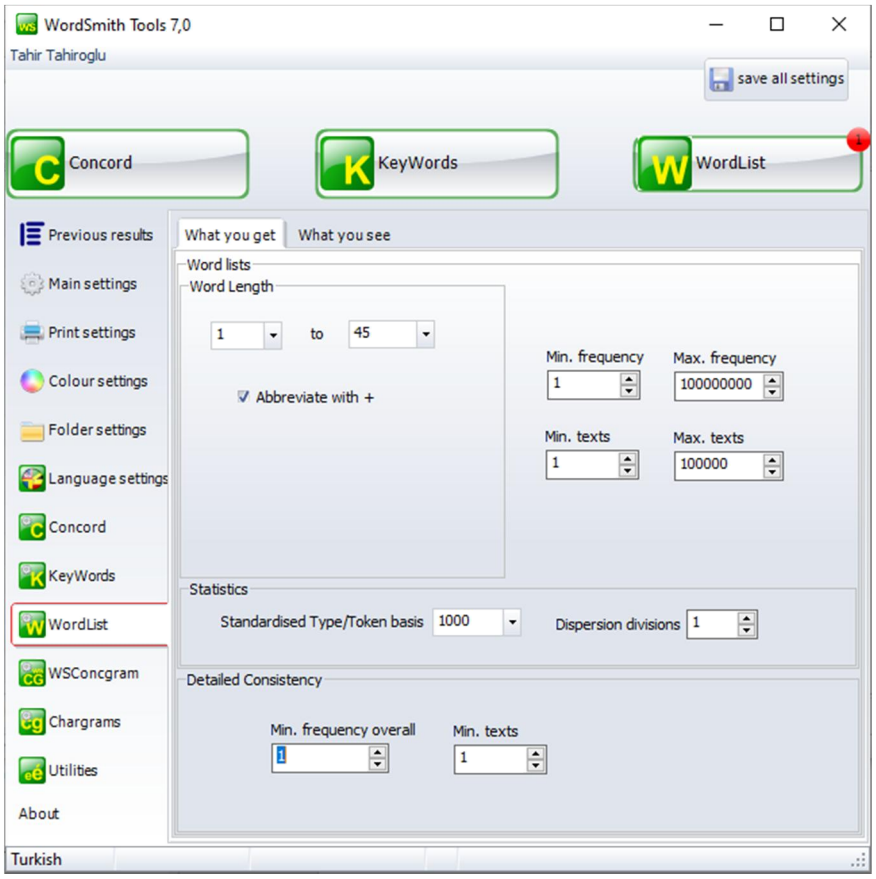
2. Yöntem

Çalışmada Muharrem ERGİN tarafından çeviriyazısı yapılan *Dede Korkut Kitabı*'nın Dresden nüshası esas alınmış ve metin sayısallaştırılarak metin işleme yazılımları için hazır duruma getirilmiştir. Eserin tam metni üzerine birçok çalışma yer almaktadır. Muharrem ERGİN'in tarafından yapılan çeviriyazılı metinin bu çalışmada tercih edilmesinin nedeni alanda yaygın kullanımıdır. Bununla birlikte, yapılan diğer çeviriyazılı metinlerle bu metnin metin madenciliği bakımından karşılaştırılması da başka bir çalışma konusu olarak ele alınabilir. Metinde optik karakter okuma kaynaklı sorunlar elle düzeltilmiş ve metin baştan sona yeniden okunarak karakter hatalarından arındırılmıştır. Metinde Vatikan nüshasına ait dipnotlar ana metinden ayrı olarak kaydedilmiştir. utf-8 karakter kodlaması ve txt formatında kaydedilen metinde uzunluk işaretlerinden kaynaklanacak metin işleme yazılımlarına bağlı sorunları dışlamak için ā, î, û karakterleri çift karakter ile (aa, ii, uu) biçimine dönüştürülmüştür. İncelemede büyük-küçük harf duyarlılığı kaldırılarak tüm birimlerin sayım seviyesi eşit tutulmuştur.

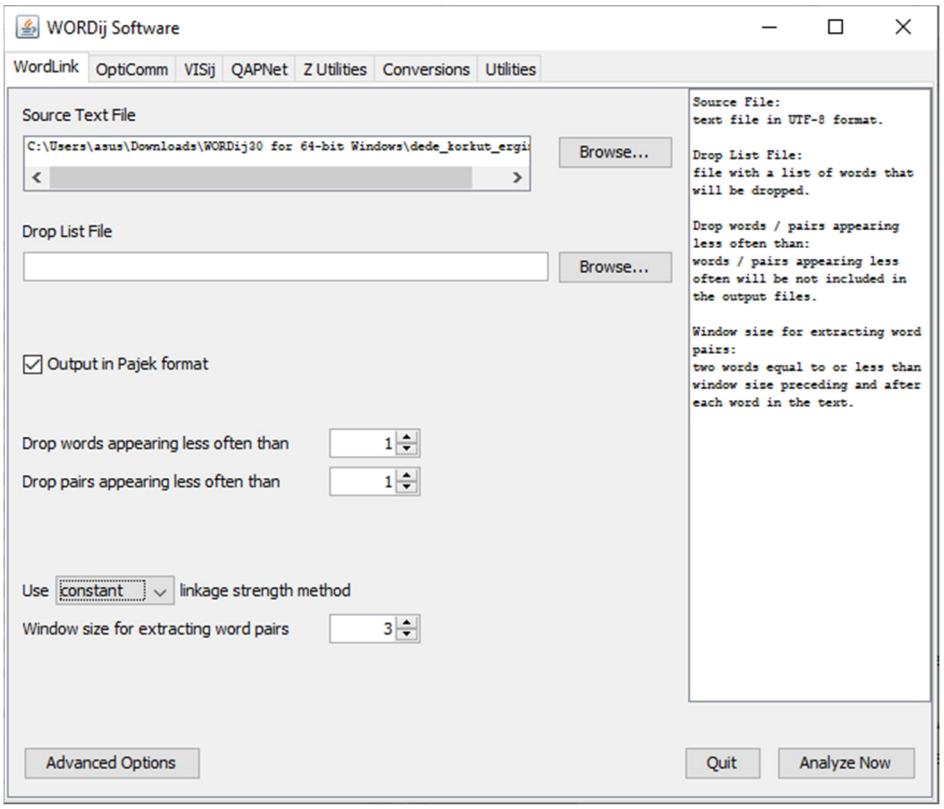
Dresden nüshası söz varlığı incelemesiyle ilgili olan bu çalışmanın yöntem açısından temel çıkış noktası nicel çözümlemedir. Metinde kullanılan noktalama işaretleri hariç her birim sayılmış, gerek frekans tablolarının oluşturulmasında gerekse sözcükler arasında ağ çözümlemesinin yapılmasında esas nokta birimleri temsil eden frekans değerleri olmuştur. Derlem dilbilimde de sık kullanılan çözümleme frekans çözümlemesi aslında sonuca uzanan aşamaların temel noktasıdır. Kategorik bir veri tipinin yazılımca değerlendirilmesi için öncelikle sözcüğün ya da birimin (token) sayısal düzlemde temsil edilmesi gerekmektedir.

Metin işleme için farklı yazılımlar kullanılmıştır. Kullanılan yazılımların metnin bütünüyle ilgili çeşitli yetenekleri söz konusudur. Kimi yazılımlar yalnızca sözcük ağı analizi için geliştirilmişken kimi metnin genel sözcük frekansları dağılımlarını gösteren araçlara sahiptir. Çalışmada metin madenciliği ve

içerik analizinde de sık kullanılan Wordstat, sözcük ağı analizi için WORDij, genel sözcük dağılımlarının hesaplanması için WordSmith adlı yazılımlar kullanılmıştır. Yazılımların sözcük frekansı alt taban seçimlerinde bir sınır konulmamış ve en düşük sözcük frekansı 1 kabul edilerek eşik düşük tutulmuş böylece tüm sözcüklerin hesaplanarak eşdizim ve ağ örüntülerinin olabildiğince güçlü olması amaçlanmıştır. Aşağıdaki şekillerde kullanılan yazılımların ekran görüntüleri verilmiştir.



Şekil 1. WordSmith 7.0 sözcük listesi oluşturma seçenekleri.



Şekil 2. WORDij sözcük bağlantısı hesaplama seçenekleri.

Java tabanlı açık kaynak kodlu bir araç olan WORDij sözcükler arasındaki anlamsal ağları hesaplamalı bir biçimde çözümleyerek ağ görselleştirmesi sağlamaktadır. Sözcük ağlarının çıkarılmasında kullanılan yöntem, kayan sözcük pencereleri (*sliding window*) yoluyla tüm metni ikili birimler (bigram) biçimindeki sözcük çiftlerinin sırayla taranmasını içermektedir. Pencerede merkeze alınan ve düğüm niteliğindeki her sözcük sağ ve solundaki diğer sözcükler arasındaki istatistiksel ilişki gücüne göre tablolaştırılmaktadır. Sözcükler arasındaki bağlantıların gücü constant, linear ve exponential olarak üç teknikle hesaplanır. WORDij her ne kadar geliştiricisi tarafından "semantic network tools" olarak adlandırılsa da doğrudan sözcüklerde kastedilen anlamları işleyememektedir. Bununla birlikte istatistiksel olarak anlamlı bir birliktelik bulunan sözcüklerin de anlamsal bakımdan birbirleriyle ilişkili oldukları düşünülmelidir. Bu çalışmada varsayılan hesaplama tekniği constant tercih edilmiştir. Diğer seçeneklerle yapılan deneylerde bu veride constant ile elde edilen bulgular arasında bir farkın olmadığı gözlenmiştir. Farklı metin verilerinden çıkarılan sözcük ağ yapılarının karşılaştırılması yazılımın diğer yetenekleri arasındadır.

Sonuçta yöntem olarak frekans temsilinden hareketle kullanılan yazılımların varsayılan sözcük işleme ayarları seçilerek niceliksel ve betimlemeli bir yol izlenmiştir.

3. Bulgular

Çalışmanın söz varlığına ait bulguları sözcük frekans bulguları, eşdizimlilik bulguları, sözcük ağlarıyla ilgili bulgular olmak üzere üç ana başlık altında gösterilebilir.

Söz varlığına ait betimsel istatistikler aşağıdaki şekilde gösterilmiştir. Şekilde yer alan standart type/token oranı her 1000 sözcükte tekrarlı birimlerin farklı birimlere oranını göstermektedir. İşlemin standardize edilmesi ölçütün değerleri temsil etme gücünü artırmaktadır.

tokens (running words) in text	32.434
tokens used for word list	32.434
sum of entries	
types (distinct words)	7.715
type/token ratio (TTR)	23,79
standardised TTR	58,59
STTR std.dev.	39,30
STTR basis	1.000
mean word length (in characters)	5,64
word length std.dev.	2,21

Şekil 3. Dede Korkut Kitabı Dresden nüshası sözcük istatistiği

Yukarıdaki şekle göre metin verisinde kullanılan toplam sözcük 32.434'tür. Bu sayı bir birimin tekrarlı sayımından oluşan tokenları ifade etmektedir. Token metinde geçen herhangi bir birimdir, bu birim noktalama işareti ya da rakam olabilir. Tokenlar tekrarlı yani frekansları toplamlı sayılan birimlerdir. Bir kez sayılan birimlere type adı verilmektedir. Buna göre bir kez sayılan farklı birim sayısı 7.715'tir. Eserde 12 hikâye düşünüldüğünde bu sayının yüksek bir sayı olduğu söylenebilir.. Bu çalışma içinde yer verilmeyen sözlükbirimleştirme işlemi sonrası bu sayının azalması söz konusu olacaktır. Bu durumda da sözcük ailelerine ait özelliklerin ortaya çıkacağı bir kavram alanı analizi yapılabilecektir. type/token oranı %23,79'dur. Daha önce de belirtildiği gibi standardize edildiğinde bu oranın da farklı sözcük sayısı göz önüne alındığında çıkan değer yüksek olarak yorumlanabilir. Standardize type/token oranı sözcük çeşitliliğinin de bir ölçütüdür. Sözcüklerin ortalama uzunluğu 5,64 standart sapması 2,21'ir. Standart sapmaya bakıldığında sözcük uzunluklarının ortalamadan sapmalarının düşük olduğu görülmektedir.

3.1. Sözcük frekansları

Dresden nüshasında WordSmith 7.0 ile elde edilen en yüksek frekansa sahip ilk 100 sözcük aşağıdaki tabloda verilmiştir. Tabloya bakıldığında ayırt edici ya da metni temsil edecek sözcüklerin ilk 10 sözcük olduğu düşünüldüğünde "didi" ve "aydur" sözcüklerinin yakın anlamlı olarak art arda bulunmaları dikkat çekicidir. Bu bulgu metnin anlatı özelliğinin bir göstergesi olarak yorumlanabilir. Aynı kavram alanına ait "soylamış" sözcüğü buna karşın 17. sırada yer almıştır. Dresden nüshasının "dimek" fiili ile temsil edildiği, ekli biçim olarak da belirli geçmiş zaman ekli "didi" biçiminin geçmişe dönük hikâyelemede kullanıldığı görülmektedir. İlk 20 sözcük içinde yer alan diğer "gel-", "ol-" filllerinin en çok kullanılan fiiller oluğu bunlarda da belirli geçmiş zaman biçiminin hikayelemeye koşut seçildiği görülmektedir.

Tablo 1. Dede Korkut Kitabı Dresden nüshası frekansa göre ilk 50 sözcük

sıra	sözcük	sıklık	%	dağılım
1	didi	562	1,73	0,89
2	aydur	520	1,60	0,90
3	bir	392	1,21	0,91
4	kara	349	1,08	0,89
5	kazan	235	0,72	0,63
6	ne	233	0,72	0,94
7	geldi	205	0,63	0,92
8	oğul	194	0,60	0,74
9	ağ	190	0,59	0,90
10	bu	183	0,56	0,88
11	hanum	174	0,54	0,93
12	oldı	162	0,50	0,83
13	ol	155	0,48	0,86
14	oğlı	153	0,47	0,87
15	dahı	149	0,46	0,86
16	mere	149	0,46	0,81
17	soylamış	148	0,46	0,76
18	manga	129	0,40	0,91
19	olsun	129	0,40	0,83
20	beyrek	127	0,39	0,40
21	oğuz	122	0,38	0,90
22	kaafir	120	0,37	0,76
23	big	119	0,37	0,79
24	menüm	117	0,36	0,84
25	yigit	115	0,35	0,83
26	kan	108	0,33	0,57
27	görelüm	106	0,33	0,88
28	delü	105	0,32	0,57
29	ala	104	0,32	0,82
30	kırk	102	0,31	0,67
31	sanga	99	0,31	0,83
32	han	96	0,30	0,61
33	kız	94	0,29	0,63
34	kim	92	0,28	0,86
35	at	91	0,28	0,82
36	men	90	0,28	0,79
37	aldı	87	0,27	0,87
38	oğlan	86	0,27	0,59

39	didiler	84	0,26	0,77
40	görklü	84	0,26	0,75
41	sen	83	0,26	0,84
42	üzerine	81	0,25	0,89
43	senüng	77	0,24	0,81
44	iki	74	0,23	0,85
45	yok	73	0,23	0,78
46	gördi	69	0,21	0,78
47	gün	69	0,21	0,80
48	bigler	66	0,20	0,72
49	var	65	0,20	0,77
50	seni	64	0,20	0,88
51	virdi	64	0,20	0,88
52	yire	62	0,19	0,88
53	üç	61	0,19	0,84
54	koca	60	0,18	0,66
55	bayındır	59	0,18	0,71
56	olur	59	0,18	0,71
57	karşu	58	0,18	0,88
58	kızı	58	0,18	0,74
59	diyü	56	0,17	0,89
60	burada	55	0,17	0,72
61	haber	54	0,17	0,82
62	altun	53	0,16	0,76
63	kibi	52	0,16	0,66
64	kalın	51	0,16	0,79
65	kazılık	51	0,16	0,70
66	tangrı	51	0,16	0,69
67	korkut	50	0,15	0,66
68	yetdi	50	0,15	0,57
69	yirde	50	0,15	0,85
70	çoban	49	0,15	0,25
71	kazanung	49	0,15	0,50
72	ozan	49	0,15	0,40
73	yidi	49	0,15	0,78
74	bigleri	48	0,15	0,75
75	böyle	48	0,15	0,71
76	dirse	48	0,15	0,00
77	pay	48	0,15	0,41
78	allah	47	0,14	0,76

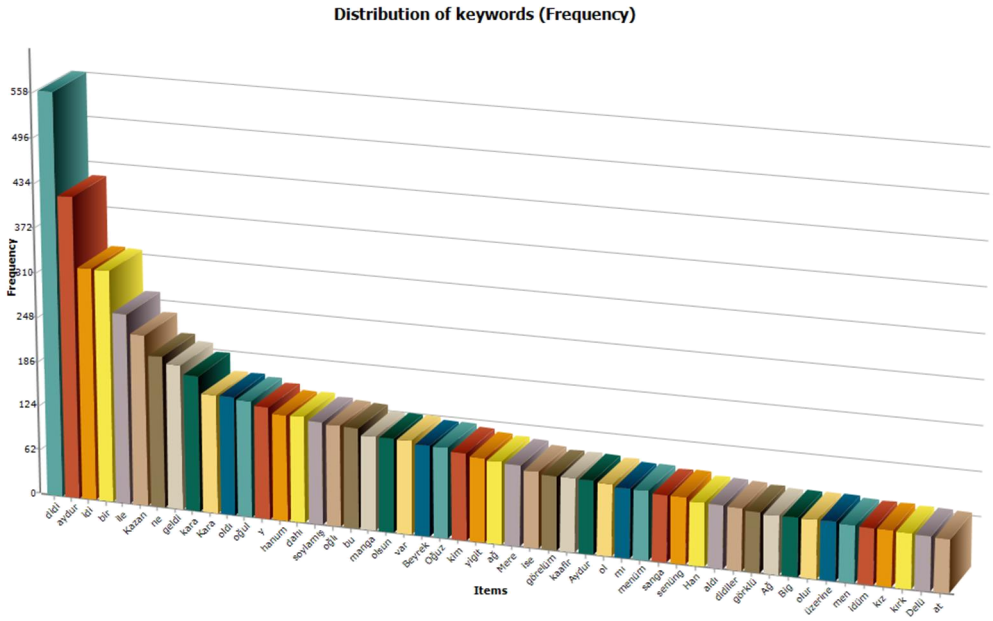
79	dede	46	0,14	0,31
80	ağam	45	0,14	0,62
81	al	45	0,14	0,78
82	er	45	0,14	0,68
83	eyledi	45	0,14	0,85
84	turalı	45	0,14	0,22
85	aruz	44	0,14	0,40
86	yir	44	0,14	0,85
87	basat	43	0,13	0,15
88	hanung	43	0,13	0,58
89	meger	43	0,13	0,84
90	altı	42	0,13	0,80
91	başum	42	0,13	0,77
92	meni	42	0,13	0,75
93	ola	42	0,13	0,83
94	sakallu	42	0,13	0,82
95	adam	40	0,12	0,66
96	depegöz	40	0,12	0,13
97	digil	40	0,12	0,69
98	bing	39	0,12	0,70
99	girü	39	0,12	0,74
100	kaba	39	0,12	0,87

Sözcük frekansları azaldıkça sözcüklerin en çok kullanılan biçiminin yerini diğer biçimlerine bırakması olasılığı artmaktadır. Aşağıdaki tabloda frekans azaldıkça sözcüğün kullanılan diğer biçimlerinin sıralanmaya başladığı ve alfabetikleşme eğilimine girdiği gözlenmektedir.

Tablo 2. Dede Korkut Kitabı Dresden nüshası azalan sözcük frekansı

sıra	**sözcük**	**sıklık**
2447	didügümi	2
2448	didügüng	2
2449	didükleri	2
2450	dikdiler	2
2451	dikdüreyim	2
2452	dikem	2
2453	dikildi	2
2454	dikilmiş	2
2455	dikmiş-idi	2
2456	dilediler	2
2457	dilek	2
2458	dilerem	2

2459	dileyeni	2
2460	dileyü	2
2461	dileyüpdür	2

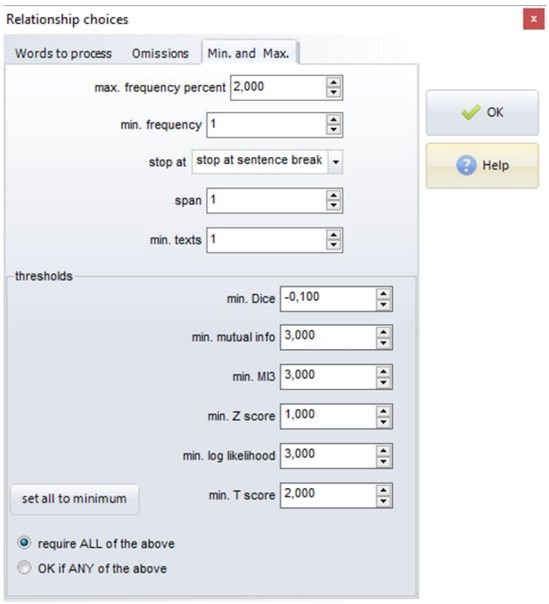


Şekil 4. WordStat 8.0 yazılımında Dresden Nüshası sıklık bar çubuğu gösterimi.

Şekil 6'da "didi", "aydur" ve "soylamış" biçimlerinin ilk 20 sözcük arasında yer alması dikkat çekicidir. Dede Korkut'ta söylemek ya da demek kavramlarının ilk bakışta önemini göstermektedir.

3.2. Eşdizimlilik

Dresden nüshasında eşdizim çıkarımında WordSmith 7.0'da relationship seçeneği kullanılmıştır. Tüm sözcüklerin oluşturulan dizini üzerinden sözcükler arası istatistiksel ilişkilerin hesaplandığı pencerenin görüntüsü aşağıda verilmiştir. Yazılımın varsayılan değer ayarları değiştirilmemiştir.



Şekil 5. WordSmith 7.0 eşdizimsel çıkarım seçenekleri

İki birimden oluşan eşdizimlilerin ele alındığı çalışmada istatistik skorlara göre farklı görünümler elde edilmiştir. Aşağıdaki tabloda iki sözcük arasındaki istatistik ilişkilerin gücüne göre uygulanan değerlerden örnekler yer almaktadır.

Tablo 3. İki birimin yan yana hesaplandığı (joint) frekans değerine göre ilk 25 eşdizimli birimler

Sıra	Sözcük 1	Sıklık	Sözcük 2	Sıklık	Joint	Log L.	T score	Dice	Log Ratio
1	aydur	520	mere	149	**88**	417,62	9,13	0,26	1,80
2	soylamış	148	görelüm	106	**61**	443,09	7,75	0,48	0,48
3	ne	233	soylamış	148	**60**	320,33	7,61	0,31	0,65
4	kan	108	turalı	45	**44**	441,32	6,61	0,58	1,26
5	oğuz	122	bigleri	48	**39**	342,03	6,22	0,46	1,35
6	ağ	190	sakallu	42	**37**	302,62	6,04	0,32	2,18
7	mere	149	kaafir	120	**30**	151,30	5,38	0,22	0,31
8	pay	48	püre	31	**29**	332,75	5,38	0,73	0,63
9	böyle	48	digeç	29	**28**	326,53	5,28	0,73	0,73
10	bayındır	59	hanung	43	**28**	266,30	5,28	0,55	0,46
11	kazan	235	big	119	**28**	111,83	5,13	0,16	0,98
12	ala	104	gözlü	34	**27**	241,99	5,18	0,39	1,61
13	aydur	520	oğul	194	**26**	38,02	4,49	0,07	1,42
14	kara	349	başum	42	**25**	137,37	4,91	0,13	3,05
15	kara	349	göne	25	**24**	177,24	4,84	0,13	3,80
16	bir	392	dahı	149	**23**	50,18	4,42	0,09	1,40
17	didi	562	beyrek	127	**23**	42,83	4,34	0,07	2,15
18	delü	105	dumrul	24	**22**	210,53	4,67	0,34	2,13

19	hanum	174	hey	37	**22**	151,51	4,65	0,21	2,23
20	başum	42	kurban	30	**20**	205,33	4,46	0,56	0,49
21	karşu	58	yatan	30	**20**	190,93	4,46	0,45	0,95
22	mere	149	kavat	35	**18**	121,90	4,20	0,20	2,09
23	delü	105	karçar	20	**17**	156,00	4,11	0,27	2,39
24	didi	562	böyle	48	**17**	54,87	3,92	0,06	3,55
25	ağ	190	pürçeklü	16	**16**	143,16	3,98	0,16	3,57

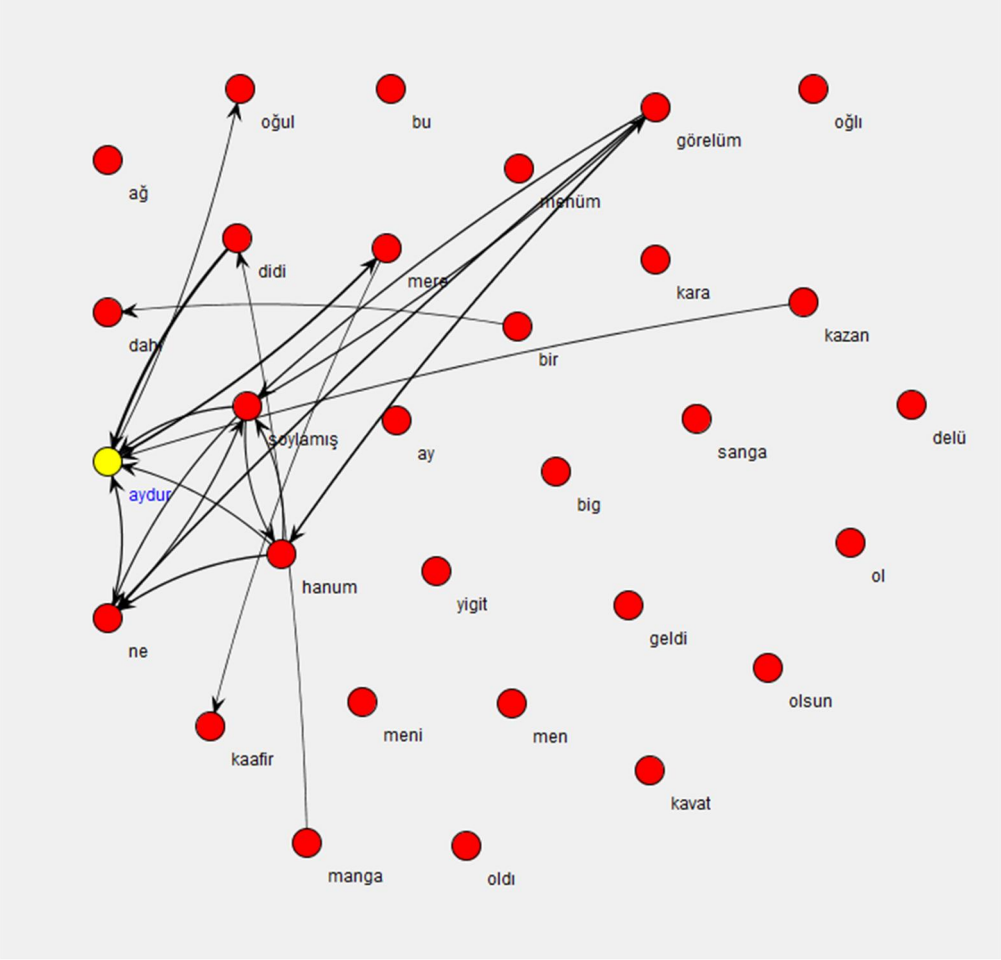
Tablo 3'te iki birimin yan yana frekanslarına göre dizilimleri yer almaktadır. Bu bir anlamda metinde birim ardıllanmasının fazladan bir formüle gerek duyulmadan hesaplanmasıdır. Burada sayılan birimler kendi değerleri üzerinden bir örüntü oluşturmakta, metindeki diğer birimler hesaplamaya dahil edilmemektedir. Frekans düştükçe birimler arasındaki anlamlı bağlılık da azalmaktadır.

Tablo 4. iki birimin log likelihood değerine göre ilk 25 eşdizimli birimler

sıra	**sözcük 1**	**sıklık**	**sözcük 2**	**sıklık**	**mı3**	**log l.**	**t score**	**dice**	**log ratio**
1	soylamış	148	görelüm	106	18,84	**443,09**	7,75	0,48	0,48
2	kan	108	turalı	45	19,12	**441,32**	6,61	0,58	1,26
3	aydur	520	mere	149	18,13	**417,62**	9,13	0,26	1,80
4	oğuz	122	bigleri	48	18,33	**342,03**	6,22	0,46	1,35
5	pay	48	püre	31	19,02	**332,75**	5,38	0,73	0,63
6	böyle	48	digeç	29	18,97	**326,53**	5,28	0,73	0,73
7	ne	233	soylamış	148	17,63	**320,33**	7,61	0,31	0,65
8	ağ	190	sakallu	42	17,65	**302,62**	6,04	0,32	2,18
9	bayındır	59	hanung	43	18,10	**266,30**	5,28	0,55	0,46
10	ala	104	gözlü	34	17,46	**241,99**	5,18	0,39	1,61
11	delü	105	dumrul	24	17,07	**210,53**	4,67	0,34	2,13
12	başum	42	kurban	30	17,65	**205,33**	4,46	0,56	0,49
13	karşu	58	yatan	30	17,19	**190,93**	4,46	0,45	0,95
14	kara	349	göne	25	15,65	**177,24**	4,84	0,13	3,80
15	muhammede	14	salavat	11	18,10	**160,24**	3,32	0,88	0,35
16	delü	105	karçar	20	16,21	**156,00**	4,11	0,27	2,39
17	bellü	19	bilgil	11	17,66	**152,04**	3,31	0,73	0,79
18	hanum	174	hey	37	15,71	**151,51**	4,65	0,21	2,23
19	mere	149	kaafir	120	15,58	**151,30**	5,38	0,22	0,31
20	berü	20	gelgil	15	17,51	**149,78**	3,46	0,69	0,42
21	kıyan	14	selçük	10	17,82	**145,19**	3,16	0,83	0,49
22	yetdi	50	çal	18	16,59	**144,94**	3,73	0,41	1,47
23	ağ	190	pürçeklü	16	15,42	**143,16**	3,98	0,16	3,57
24	kara	349	başum	42	15,08	**137,37**	4,91	0,13	3,05
25	sorar	24	olsam	19	16,91	**134,69**	3,46	0,56	0,34

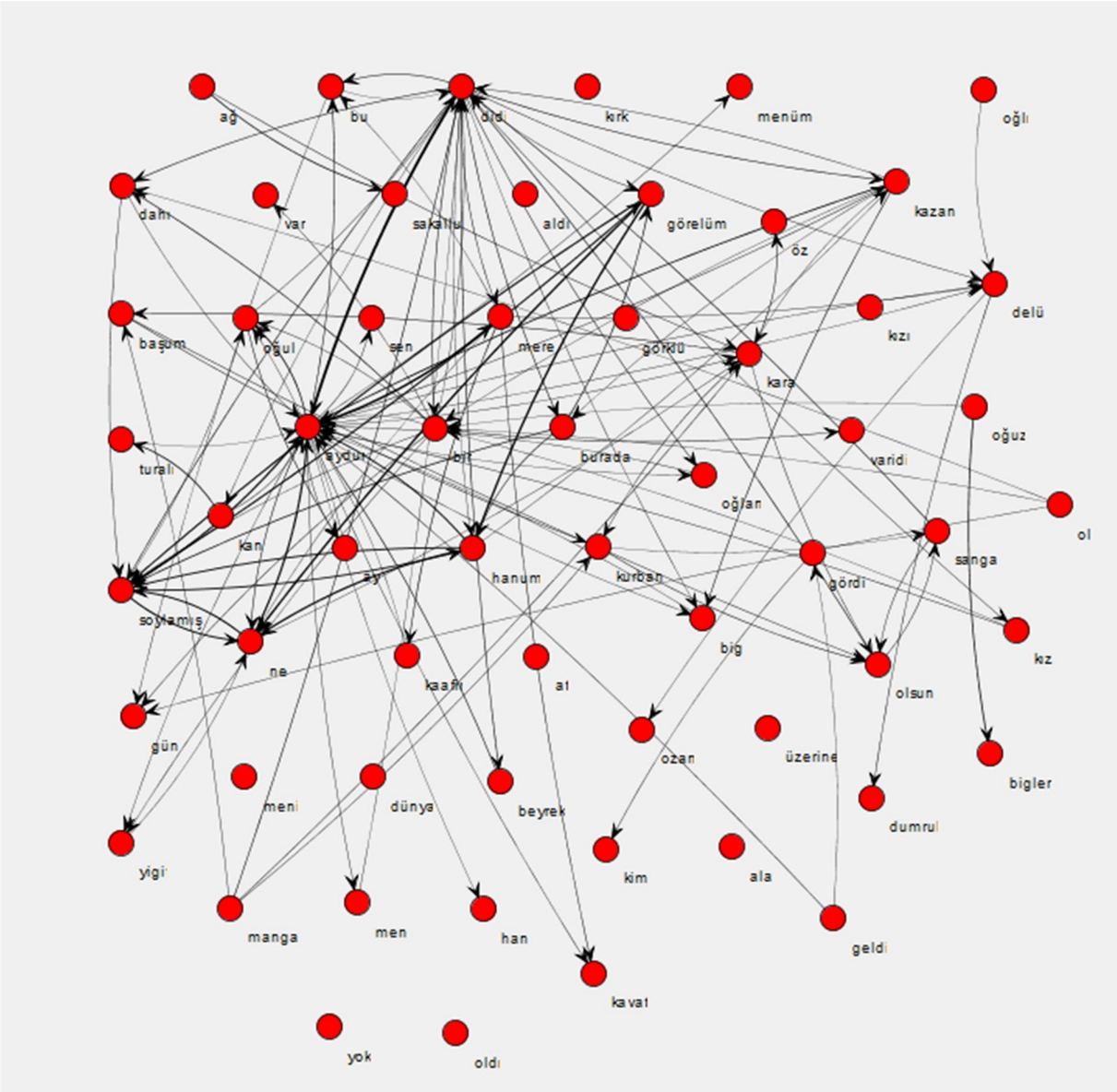
3.3 Sözcük ağları

Metin madenciliği yazılımlarının sağladığı olanaklardan biri olan sözcükler arası bağlantı (link) analizinde, metinde ya da ilgilenilen derlemde bulunan tüm sözcüklerin birbirleriyle benzerliklerine, bir arada bulunabilirliklerine göre hesaplanmasıyla çıkarılan ağlar (network) görselleştirilmektedir. Bu çalışmada kullanılan WordStat ve WORDij yazılımlarıyla elde edilen ve Dresden nüshasındaki tüm sözcüklerin hesaplanmasıyla oluşturulan ağ görünümleri aşağıdaki şekillerde yer almaktadır.



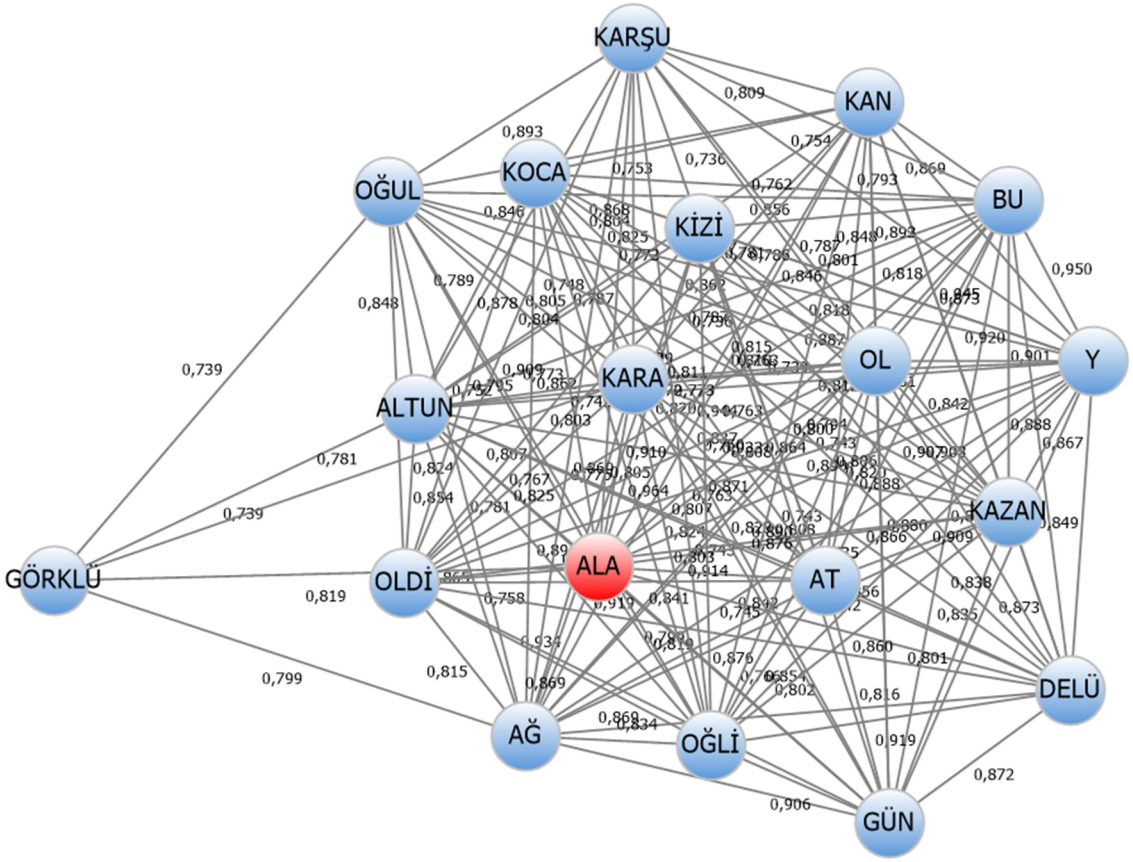
Şekil 6. WORDij yazılımıyla Dresden nüshasında ağ yapan 30 düğüm sözcük gösterimi

Şekil 6'de oluşturulan ağ görüntüsünün hesaplanmasında 30 düğüm sözcük kullanılmıştır. Sözcükler arasındaki bağlantıların gücünün hesaplanması için girilen eşik değer de 30 olarak belirlenerek en güçlü ilişki kuran sözcüklerin belirginleştirilmesi sağlanmıştır. Şekilde daha önce sıklık tablosunda da gösterilen "aydur", "soylamış" ve "didi" biçimleri arasındaki ilişkinin daha çok "aydur" yönünde bağlantı aldığı görülmektedir. "didi"ye gelen bağlantı sayısı az, "didi"den "aydur"a çıkan bağlantı gücü daha fazladır. Düğüm sözcük sayısı 60'a çıkarılıp bağlantı gücü eşik değeri 10 olarak değiştirildiğinde aşağıda şekil 7'deki ağ görüntüsü oluşmaktadır. Buna göre metnin ana kavram ağlarının yine sözü edilen üç sözcük etrafında oluştuğu bunlara ek olarak "kara", "kazan", "görelüm", "big" ve "olsun" "delü" biçimlerinin de etraflarında ağlar oluşturduğu gözlenmektedir.



Şekil 7. WORDij yazılımıyla Dresden nüshasında ağ yapan 60 düğüm sözcük gösterimi.

WordStat 8.0 ile yapılan sözcükler arası link analizinde aşağıdaki görselleştirme elde edilmiştir. Analiz sırasında ilişkilerin bulunması için aynı paragrafta yer alma ve sözcüklerin ikili sıralanım düzeni (ngram =2) kullanılmıştır. Şekilde sözcükler arasındaki bağlantı gücü doğrudan puanlarla gösterilmiştir. Ağ görünümünde gösterilen bu oluşumlar aslında bir çeşit eşdizimlilik gösterimi kısaca birlikte bulunabilirliklerin gösterimidir. Buna göre paragraf düzeyinde bakıldığında Dresden nüshasında "karşu", "koca", "delü", "oğlı", "görklü", "kazan" sözcüklerinin ön plana çıktığı görülmektedir. WORDij ile tüm metnin işlendiği sonuçlardaki fiil sözcüklerin yerini paragraf düzeyinde isim, sıfat ve zamir görevli sözcüklerin aldığı görülmektedir.



Şekil 8. WordStat 8.0 yazılımıyla Dresden nüshası paragraf düzeyinde sözcük ağları bulgusu.

Sonuç

Dede Korkut Kitabı'nın Dresden nüshasının metin madenciliğinde kullanılan bazı yazılımlarla ele alındığı bu çalışmada elde edilen bulgulardan Dresden nüshasının söz varlığına dair görünümlerin metnin farklı bir açıdan okumasının yapılmasına olanak sağlayacağı sonucuna varılabilir. Özellikle ilk 100 sözcük sıklığı tablosundan metnin sıklık temelli anahtar sözcükleri olarak *di-*, *ayt-* fiilleriyle oğul adının ön plana çıktığı görülmektedir. Eşdizimlilik örüntülerinde de ilk 25 arasında "kan turalı", "oğuz bigleri" gibi özel adların yanında; "ağ pürçeklü" ve "ala gözlü", "ağ sakallu" gibi niteleyici sözlerin istatistiksel olarak belirgin olduğu görülmüştür. İlk 100 sözcük sıklığının ilk 30 sözcüklük diliminde fiilerin ağırlık göstermesi metnin hikâye tarzı anlatımının kanıtlayıcı özelikleri durumundadır. Niteleyici ifadelerin eşdizimli sözcüklerde görülmesi de hesaplamalı tekniklerin ilk bakışta görünmeyen yapıların bulunmasındaki rolünü göstermektedir.

Sözcük sıklıkları bakımından elde edilen sonuçlar bakımından Çitgez 2018'de yer alan sonuçlar karşılaştırıldığında benzer sonuçlar görülmektedir. İlk 20 sözcüğün sıklığı aşağıdaki şekillerde gösterilmiştir.

sıra	sözcük	sıklık	%	dağılım
1	didi	562	1,73	0,89
2	aydur	520	1,60	0,90
3	bir	392	1,21	0,91
4	kara	349	1,08	0,89
5	kazan	235	0,72	0,63
6	ne	233	0,72	0,94
7	geldi	205	0,63	0,92
8	oğul	194	0,60	0,74
9	ağ	190	0,59	0,90
10	bu	183	0,56	0,88
11	hanum	174	0,54	0,93
12	oldı	162	0,50	0,83
13	ol	155	0,48	0,86
14	oğlı	153	0,47	0,87
15	dahı	149	0,46	0,86
16	mere	149	0,46	0,81
17	soylamış	148	0,46	0,76
18	manga	129	0,40	0,91
19	olsun	129	0,40	0,83
20	beyrek	127	0,39	0,40

Şekil 9. Dersden nüshası ilk 20 sözcük sıklığı.

SIRA NO	DEDE KORKUT HİKÂYELERİ'NDE EN ÇOK KULLANILAN 20 SÖZCÜK	KULLANIM SAYISI
1	***di-***	973
2	***ol-***	695
3	***ayıt-***	589
4	***gel-***	545
5	*i-*	495
6	***bir***	439
7	*men*	405
8	*gör-*	388
9	*sen*	372
10	*kara*	371
11	*han*	353
12	*oğul*	348
13	*kazan*	341
14	*beg*	337
15	*yir*	310
16	*ne*	304
17	*at*	303
18	***bu***	294
19	***vir-***	290
20	***ol***	286

Şekil 10. Çitgez 2018'de ilk 20 sözcük sıklığı. (Çitgez, 2018, s. 28)

Şekillerde *di-* ve *ayt-* fiillerinin iki çalışmada da birincil söz varlığı ögeleri olduğu görülmektedir. Çitgez 2018'de sözcüklerin bizim çalışmamızda olduğu gibi çekimli biçimleri değil maddebaşı biçimlerinin sayımı esas alındığından sıralama farklılıkları görülmektedir. Bununla birlikte iki çalışmanın da sözcük sıklıkları açısından benzer oldukları söylenebilir.

Metin madenciliği ve tekniklerinin Türkçenin tarihsel metinlerinde şimdiye kadar yapılmamış farklı çalışmalarda kullanılabilmesi hem kültürel özelliklerin söz varlığındaki izlerinin ortaya çıkarılmasına hem de tarihsel sözlüklerin hazırlanması ve geliştirilmesine büyük katkılar sağlayacaktır.

Kaynakça

Akbıyık, A. (2019). *Sosyal Bilimlerde Metin Madenciliği*. Sakarya: Sakarya.

Aksan, D. (2018). *Türkçenin Sözvarlığı* (2nd ed.). Ankara: Bilgi.

Altunkaynak, B. (2019). *Veri Madenciliği Yöntemleri ve R Uygulamaları* (2. baskı). Ankara: Seçkin.

Anandarajan, M., Hill, C., ve Nolan, T. (2019). *Practical Text Analytics: Maximizing the Value of Text Data*. *Advances in Analytics and Data Science: Vol. 2*. Cham: Springer International.

Çitgez, M. (2018). *Dede Korkut Hikâyeleri'nin Söz Varlığı,* Basılmamış Doktora tezi*,* T.C. Ardahan Üniversitesi Sosyal Bilimler Enstitüsü

Danowski, J. A. (2013). WORDij version 3.0: Semantic network analysis software. Chicago: University of Illinois at Chicago.Ergin, M. (1994). *Dede Korkut Kitabı I*. Ankara: Türk Dil Kurumu.

Günay, D. (2018). *Sözcükbilime Giriş* (2. baskı). İstanbul: Papatya.

Gürsoy, U. T. Ş. (2009). *Veri Madenciliği ve Bilgi Keşfi*. Ankara: Pegem Akademi.

İmer, K., Kocaman, A., & Özsoy, A. S. (2011). *Dilbilim sözlüğü* (1. basım). Etiler İstanbul: Boğaziçi Üniversitesi.

Karaağaç, G. (2013). *Dil bilimi terimleri sözlüğü* (Birinci baskı: Ankara, 2013 Şubat). *Atatürk Kültür, Dil ve Tarih Yüksek Kurumu Türk Dil Kurumu Yayınları: 1066*. Ankara: Türk Dil Kurumu.

Korkmaz, Z. (1998). Dede Korkut Hikayelerinde Dil-Üslup Bağlantısı. *TDAY Belleten*, *46*, 101–112.

Oğuzlar, A. (2011). *Temel Metin Madenciliği*. Bursa: DORA Basım-Yayın Dağıtım.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri*. İstanbul: Papatya. Scott, M. (2016). WordSmith Tools version 7, Stroud: Lexical Analysis Software.

Silahtaroğlu, G. (2008). *Kavram ve Algoritmalarıyla Temel Veri Madenciliği*. İstanbul: Papatya.

Vardar, B. (1998). *Açıklamalı Dilbilim Terimleri Sözlüğü*. İstanbul: ABC.

Elektronik kaynaklar

WordStat 8.0, https://provalisresearch.com/products/content-analysis-software/, (Erişim tarihi: 14.02.2021)