TÜRKÇENİN DERLEM-TEMELLİ SIKLIK SÖZLÜĞÜ: TEMEL İLKELER VE UYGULAMA

A CORPUS-BASED FREQUENCY DICTIONARY OF TURKISH: BASIC PRINCIPLES AND IMPLEMENTATION[3]

Yeşim Aksan; Mustafa Aksan

**Mersin Üniversitesi, İngiliz Dili ve Edebiyatı Bölümü, e-posta: yaksan@mersin.edu.tr*

*** Mersin Üniversitesi, İngiliz Dili ve Edebiyatı Bölümü, e-posta: maksan@mersin.edu.tr*

Özet

Geçerliliği ve güvenirliği sınanmış büyük bir dil derlemi olan Türkçe Ulusal Derlemi üzerinden bir sözcük ve ek sıklığı çalışması yeni tamamlanmış ve bu çalışmanın ürünü olarak Türkçe Sözcük ve Ek Sıklığı Sözlüğü basıma hazır hale getirilmiştir. Bu çalışmanın amacı derlem dilbilim ve sayısal dilbilim yöntemleri kullanılarak hazırlanan sözkonusu sözlüğü oluşturmada izlenen tasarım ilkelerini ve uygulamaları anlatarak sözlüğü kapsam ve içerik olarak betimlemektir. Ayrıca bu sözlüğün hazırlanmasında Türkçenin çok sözcüklü birimlerinin (ÇSB)lerin belirlenmesi ve sözlüğünün hazırlanmasına ilişkin yapılan saptamalar da yer verilmiştir.

Anahtar Sözcükler: Türkçe Ulusal Derlemi (TUD), sözcük sıklığı, ek sıklığı, çok sözcüklü birim

---

3 Bu çalışma 113K039 ve 115K135 numaralı TÜBİTAK (SOBAG-1001) projeleri kapsamında yapılmıştır. Katkılarından dolayı TÜBİTAK'a teşekkür ederiz.

Abstract

A recently completed study on word frequencies in Turkish, extracted from a large-scale, reliable and representative corpus of Turkish National Corpus, has produced A Corpus-based Frequency Dictionary of Turkish, listing lexical items and affixes. The aim of this study is to present basic principles adopted and implemented in preparing this dictionary and to introduce its coverage and content. This research is also noted that the processes and principles followed in preparing the frequency dictionary have provided insights for proposing a new project aiming at making a corpus-based multi-word unit dictionary of Turkish.

Keywords: *Turkish National Corpus (TNC), word frequency, suffix frequency, multi-word units*

1. Giriş

Dilin temel birimlerinin kullanım sıklıkları, sayısal tanımlamaya bağlı olarak dilin yapısal tercihlerinin oluşumu ve değişimi, son dönem çalışmalarda ilgi çeken bir alan olmuştur. Uzun bir geçmişi olan dilde sıklık çalışmaları, önceki dönemlerde daha çok uygulama alanlarına dönük olarak gelişmişse de, özellikle derlem dilbilimi alanındaki gelişmelere bağlı olarak, daha kuramsal boyutta taşınmıştır. Günümüzde, dil birimlerinin sıklıkları, kullanım ortamları, dağılım özellikleri, dil çözümlemelerinin ayrılmaz bir parçası olmaya doğru ilerlemektedir.

Geçerliliği ve güvenirliği sınanmış, büyük bir dil derlemi olan Türkçe Ulusal Derlemi (TUD) (www.tnc.org.tr) temellinde bir sözcük ve ek sıklığı çalışması yeni tamamlanmış ve bu çalışmanın ürünü olarak *Türkçe Sözcük ve Ek Sıklığı Sözlüğü* basıma hazır hale getirilmiştir. Bu çalışmanın amacı derlem dilbilim ve sayısal dilbilim yöntemleri kullanılarak hazırlanan sözkonusu sözlüğü oluşturmada izlenen tasarım ilkelerini ve uygulamaları anlatarak sözlüğü kapsam ve içerik olarak betimlemektir. Ayrıca, bu sözlüğün hazırlanmasında Türkçenin çok sözcüklü birimlerinin (ÇSB) lerin belirlenmesine ilişkin yapılan saptamaları da, çalışma kapsamında tartışmaya açmaktır.

2. Türkçe Sıklık Çalışmaları

2.1. Öncü Çalışmalar

Türkçe için yapılan sıklık çalışmalarının başlangıcı 1960'lara kadar tarihlenebilse de, henüz bütüncül, kapsamlı, tutarlı, nitel ve nicel olarak güncel Türkçenin tüm *başsözcük* (İng. lemma) ve biçimbirimlerini içeren, sözcük türü işaretlemelerine dayalı bir sıklık sözlüğü yayınlanamamıştır.

Öncü çalışmalar arasında, Türkçe biçimbirim sıklarıyla ilgili araştırmalar yapan Pierce (1961, 1962)'yi gösterebiliriz. Pierce (1961), %75'i okuma-yazma bilmeyen fabrika işçilerinin konuşmalarını bir berber dükkânında kaydetmiş (47.000 sözcük), bunlara askerlik görevini yapmakta olan, yine okuma-yazma bilmeyen erlerin konuşmalarını (93.000 sözcük) ekleyerek 140.000 sözcüklük bir sözlü derlem oluşturmuş ve yazıya aktarılmasını sağlamıştır. Listelediği toplam çekim ve türetim biçimbirimi sayısı 110'dur. Bu çalışma, Türkçede her 5 sözcükten 4'ünün çekimli ya da türemiş olduğunu belirtmesi gibi istatistik bilgiler içermesi bakımından ilginçtir. Çalışmanın en sık bulunan 21 ek olarak saptadığı eklerin tümü çekim ekidir.

Pierce (1962) ise Türkçe biçimbirimlerin yazılı metinlerden oluşmuş bir örneklem üzerinden kullanım sıklıklarını belirlemek için yapılmış bir çalışmadır. Yazılı metinler kümesinin romanlar, orduya ait saha el kitapları, devlet okullarında okutulan ders kitapları, şiirler, dini hikâyeler, kısa hikâyeler ile gazete ve dergilerden seçilmiş bazı makalelerden oluştuğu belirtilmiştir. Toplam küme 2.000.000 sözcükten oluşmakta, çalışma ise bu küme içinden alınan yaklaşık 100.000 sözcüklük bir örneklemi kullanmaktadır. Çalışmanın bulguları yazılı metin örnekleminde toplam sayısı 139 adet olan farklı biçimbirim bulunduğunu ve bunların örneklemde 60.038 defa geçtiğini göstermektedir. Buna göre, 139 birimden yalnızca 36 adedi çekim biçimbirimi olmasına rağmen, çekim biçimbirimlerinin toplam sıklığının % 74'ünü temsil etmektedir. En sık görülen 29 biçimbirim, toplam sıklığın % 78'ini temsil etmektedir. Pierce her iki çalışma arasındaki en çarpıcı farkın yazılı dil örnekleminde bulunan biçimbirimlerin daha fazla çeşitlilik göstermesi olduğunu belirtir.

2.2. Derlem-Temelli Sözcük Sıklığı Çalışmaları

Türkçede Göz (2003)'ün hazırladığı sözcük sıklığı sözlüğünün ilk sözlük olduğunu görürüz. Bu sözlük Brown Derlemi örnek alınarak hazırlanan 1 milyon sözcüklük genel Türkçenin kullanımlarını yansıtan 1995-2000 yıllarında yayınlanan kitap ve süreli yayınlar kullanılarak yapılmıştır. Sözlük 22.693 başsözcüğün gözlenen sıklık değerlerinin sayısal ve alfabetik sıralı sıklık listelerinden oluşmaktadır. Sözcük sayısı küçük bir derlemle çalışan Göz'ün bu sözlüğünün önemli yanı çokanlamlı ve eşyazımlı başsözcüklerin anlamına göre bir sıklık değeri de verilmiş olmasıdır.

Göz (2003)'ün derlem kurma ilkelerini izleyerek Ölker (2011) 1945-1950 arası yazılı Türkçenin sözcük sıklığı sözlüğünü hazırlamış ve sıklık sonuçlarını Göz (2003) ile karşılaştırarak elli yıllık bir zaman diliminde Türkçenin sözvarlığı ve sözcük sıklığı temelinde geçirmiş olduğu değişimi yazılı kaynaklardan yola çıkarak ortaya koymuş ve böylece artzamanlı bir sıklık çalışması yapmıştır..

Aksan ve Yaldır (2011)'de, *Türkçe Kurgusal Metinler Derlemi* (1 milyon sözcük) ve *Türkçe Süreli Yayınlar Derlemi* (1 milyon sözcük) üzerinden kesitlere (İng. register) özel sözcük sıklığı listeleri oluşturarak, bu listeler üzerinden *başsözcük/ sözcükbirim* oranı belirlenmiştir. Ayrıca iki kesitte en sık kullanılan sözcükler ve sözcük türleri de karşılaştırılmıştır.

Aksan, Yaldır ve Mersinli (2011) özel amaçlı hazırlanan *Türkçe Ders Kitapları Derlemi*'nden elde edilen sözcük sıklığı listelerini *Türkçe Ulusal Derlemi* (TUD) örnekleminden oluşturulan 250 bin sözcüklük bir genel derlem ile karşılaştırarak, Türkçe ders kitaplarındaki sözvarlığının genel dil kullanımını ne kadar yansıttığını belirlemeye çalışmışlardır. Bu araştırmalarda hem otomatik olarak hazırlanan hem de araştırmacıların elle yaptığı düzeltmelerle elde edilen sözcük listeleri kullanılmıştır.

Güngör (2003), Kumova vd. (2006) ise tümüyle bilgisayar aracılığıyla yapılmış sayımları örneklemektedir. Güngör (2003) sözcük türü ve ek bilgilerinin sayısal görünümlerine ilişin bilgiler de içermektedir. Kumova vd. (2006) dil modelleme ve metnin bilgi miktarını belirlemede Zipf yasalarını kullanarak sözcük sayısı-sözcük dağarcığı ilişkisini saptamaya çalışmışlardır.

Ancak bütünüyle bilgisayar aracılığıyla yapılan bağımlı-bağımsız biçimbirim sayımlarının, araştırmacı eliyle iyileştirildiği ve güncellendiği çalışmalara ihtiyaç vardır ve bu anlamda bizim çalışmamızda izlediğimiz melez yaklaşım geçerliliği ve güvenirliği yüksek sıklık listelerinin oluşmasını sağlamıştır.

3. Kavramsal Çerçeve ve Yöntem

Çalışmanın genel kavramsal çerçevesinde sıklık çalışmalarının geliştirdiği ve bir bölümü aşağıda anılan temel ilkelerin yanı sıra, bu çalışmada özel olarak etkili olan *Routledge Sıklık Sözlükleri Dizisi* hazırlamada bilim kurulunun tanımladığı ilkeler geçerli olmuştur. Sıklık listelerinin hazırlanmasında, tüm diller için kabul gören en iyi uygulamaların (İng. best practices) ilkeleri her aşamada gözetilmiştir.

3.1. Sıklık Sözlüğü Hazırlamada Temel İlkeler

Derlem-temelli sıklık sözlüğü hazırlamada genel olarak benimsenen temel ilkeler, sözlük sıralamasının dildeki gerçek sıralamaya mümkün olduğunca yakın olmasını hedefler. Buna göre, i. Kullanılan derlemin dil temsil yeterliliği ve dengesi (*Örneklem tanımı*, *ağırlıklandırma* (İng. weighting); ii. Gelişmiş sıklık ölçüleri (*Dağılım* (İng. dispersion); iii. Metinlerin ve derlem araçlarının doğruluğu (*Başsözcük belirleme ve sözcük türü işaretleme)* geçerli ve güvenilir bir sıklık verisi elde etmek için en öncelikli olarak başvurulacak ölçütler olarak tanımlanır.

i. Derlem: Bu çalışmada tanıtılan sıklık sözlüğü verisinin temelini *Türkçe Ulusal Derlemi* (TUD) oluşturmaktadır. Bu derlem dil temsiliyeti açısından dil kesitlerinde, zamana yayılmada çeşitlilik (9 alan yazlı bölümde, 2 alan sözlü bölümde, 23 yıla yayılan zaman, çok çeşitli yayın ortamları ve dil kullanım türleri) sağlayan; derlemde yer alan dil kesitlerinin gerçek dil kullanımına uygun dağılımını karşılayan dengeli bir oranlamaya sahip; % 98 yazılı (4974 belge) , % 2 sözlü (434 belge) toplam 5408 farklı belge, 50,997,016 sözcük ile 1990-2013 yıllarını temsil eden, Türkçenin ilk geniş kapsamlı referans derlemdir (bkz. Tablo 1).

Tablo 1. Türkçe Ulusal Derlemi Genel Sayısal Görünüm

TUD Bölümler	Sözcükbirim	Teksözcük	Başsözcük
Yazılı	49,983,288	1,316,462	71,437
Sözlü	1,013,728	114,044	13,429

ii. Gelişmiş Sıklık Ölçüleri: Gözlenen sıklık değerleri sözcüklerin derlem metinleri arasındaki dağılımını gösteremediği için tek bir metinde sık kullanılan bir sözcüğün sayısal sıklığı yüksek ve listelerdeki sırası da üstte olacaktır bu da sözkonusu listenin gerçek dil kullanımını yansıtamadığı anlamına gelecektir. Gelişim sıklık ölçüleri dağılım ölçüleri ve bunlara göre düzenlenmiş sıklık verilerini gösterir. Dağılımı gösteren tekil katsayı (Gries, 2008) ve dağılım puanına göre gözlenen sıklık değerlerinin düzenlenmesi, parça-temelli (derlemi oluşturan belgeler arası hesaplama) ya da uzaklık-temelli (derlemi oluşturan belgelerin yeniden düzenlemesine duyarlı hesaplama) yaklaşımlarla yapılmaktadır. TUD temelli sözcük sıklığı listelerini parça-temelli Julliand D dağılım indeksi'ni kullanarak oluşturduk. 0.00-1.00 arası puan içeren bu indekse göre, 1.00 puan sözcüğün derlemde eşit dağıldığını, .10 ise kimi bölümlerde sözcüğün ortaya çıktığını, derlemin birçok bölümünde hiç olmadığını ya da çok az olduğun gösterir. Dağılım puanı ile gözlenen sıklığın çarpılmasıyla elde edilen katsayı değerine göre sayısal sırası belirlenin sözcük sıklığı listelerinde Tablo 2'de örneklendiği gibi, *devlet* sözcüğü (63.907) yüksek sıklık değerine ancak düşük dağılma puanına sahip görünmektedir (0.86); öte yandan, *bile* sözcüğü (59.692) düşük sıklık değerine ama derlem içinde daha iyi bir dağılıma sahiptir (0.92) ve yüksek katsayı puanı ile (55467) de ilk üç sözcük arasında ilk sırada yer almaktadır.

Tablo 2. Gözlenen Sıklık ve Dağılım Puanı

Sıra	Başsözcük	Sözcük Türü	Gözlenen Sıklık	Dağılım Puanı	Katsayı
119	bile	ilgeç	59692	0.929236	55467
120	devlet	ad	63904	0.864883	55269
121	nasıl	belirteç	58584	0.935785	54822

iii. Metinlerin ve derlem araçlarının doğruluğu: Doğal dil verilerini bilgisayar aracılığıyla otomatik olarak, dilin çeşitli düzeylerinde, bir derlem üzerinden işlemleyerek, araştırmacı eliyle edinilemeyecek betimleyici bilgiler ya da işlevsel araçlar sunan çalışmaların tümüne *Doğal Dil İşleme* (DDİ) denmektedir. Sözcükleri ya da sözcük öbeklerini biçimbilimsel olarak ayrıştırarak çözümlemek ve ilgili dilbilimsel *işaretleri* (İng. tag; annotation), bilgisayar aracılığıyla ilgili sözcüğe atamak ise *biçimbirimsel işaretleme* (İng. morphological analysis) olarak adlandırılır. Çoğunlukla, otomatik işaretleme sonrasında, işaretlerin doğruluğunun araştırmacı eliyle düzeltilmesi ve iyileştirilmesi gerekir çünkü bir sözcüğe birden fazla işaret atanabilir ya da işaret atanamayan sözcükler olabilir.

Derlem işleme araçlarınca betimlenecek olan, bu amaçla birbirinden ayrılan ve derlem büyüklüğünü belirtmekte de kullanılan temel derlem işleme birimi *sözcükbirim* (İng. token) olarak adlandırılır. Derlemde tekrar eden sözcükbirimlerin her biri ya da derlemi oluşturan ve birbirinden farklı olan her sözcük biçimi *teksözcük* (İng. type)dür. *Başsözcük* (İng. lemma) ise derlemdeki teksözcüklerin içerdiği, sözlükte madde başı olabilecek bağımsız morfemlerin her biri ya da bir teksözcüğün çekim eklerinden arındırılmış, sözlük girdisi olabilecek yalın halidir. Sıklık sözlüğü verisinin çıkarıldığı TUD ön-işleme araçları ile işlenmiştir. Sözcükbirimleştirme ve tümcelere ayırma (Aksan, Aksan, Özel vd., 2015; Aksan, Özel, Bektaş vd., 2015) işlemleri tamamlanmış, sözcük türü açımlanmış, biçimbirim çözümlemesi ve işaretlemesi yapılmıştır. Teksözcüklerin biçimbirim işaretlenmesi Nooj-TR (Mersinli, Aksan, 2011) kullanılarak yapılmış, ayrıca elle denetim ve işaretleme ile yapay belirsizliklerden arındırılmış (örn: *yelken*; *yel+ken* değil) ve listelerde sözcük türü açısından otomatik olarak tanınmayan sözcüklerin işaretlenmesi tamamlanmıştır. Tüm bu işlemleler sonucunda kapsamlı ve güvenilir bir TUD-DDİ Sözlüğü oluşturulmuştur.

4. Bulgular

TUD-DDİ sözlüğünün 5408 derlem belgesi üzerinde işaretlediği teksözcük ve başsözcük listeleri sayısal sıralı olarak hazırlanmıştır. *Routledge Sıklık Sözlükleri* dizisinden Türkçe öğrenen öğrenciler için hazırlanan, 2016 yılında yayınlanacak olan *A Frequency Dictionary of Turkish*'de bulunan sıklık in-

dekslerinin kapsamı ve hazırlama ilkeleri aşağıda sıralanmaktadır. Bu ilkeler farklı dil ailelerine üye dillerin sıklık sözlüklerinin hazırlanmasında da uygulanmıştır (örn. Leech, Rayson ve Wilson, 2001; Davies ve Gardener, 2010 ; Čermák ve Křen, 2011).

- Listeler TUD (yazılı-sözlü) belirlenen 84,866 başsözcüğün sıklıkları temelinde oluşturulmuştur.
- Listeler sadece tekil sözcüklerden oluşur. Listelere konu olan sözcüklerin eşdizimli olduğu ögeler ya da çok sözcüklü birimler (İng. multi-word units) kapsam dışıdır.
- Sözcüklerin farklı anlamlarının sıklıkları listelerde yer almaz.
- Eşyazımdan kaynaklı doğal belirsizlikler başsözcüklerin sunumunda korunmuştur. Bu girdilerin sıklıkları ilgili sözcüklerin olası görünümlerinin sıklığı olarak okunmalıdır. Buna göre, örneğin, "at" girdisiyle ilgili tüm sıklık değerleri, hem ad hem de eylem olarak, tüm yalın ya da çekimli biçimlerinin toplam sıklığıdır.
- Özel isimler, kısaltmalar, yansıma sözcükler, sayılar ve yabancı sözcükler liste dışı tutulmuştur.
- Adcıl ve eylemcil çekim ekleri ve ek dizileri gözlenen sıklık değerlerine göre listelenmiştir.

Tablo 3 *Routledge Sıklık Sözlükleri* dizi editörlerinin belirlediği ölçütleri de temel alan sözlüğün genel özelliklerini göstermektedir.

Tablo 3. Sözlüğün Genel Özellikleri

Hedef kitle	Dil öğrenenler
Girdi sayısı	5,000
Girdi seçimi"	Juliand D dağılım puanına göre düzenlenmiş sıklık indekseleri
Temel derlem	49,983,288 TUD- yazılı 1,013,728 TUD-sözlü
Kapsanan alanlar	Bilgilendirici, kurgusal TUD-yazılı; doğal ortam ve bağlam-bağımlı TUD-sözlü
Ek bilgi	Çeviri denklikleri, örnek tümceler ve çevirileri
Ek listeler	Adcıl ve eylemcil biçimbirim listeleri, konu alanına özgü 20 tematik sözcük listesi

5. YORUM ve TARTIŞMALAR

Sözlüğe konu olan sıklık listeleri incelendiğinde, iki sorun öncelikli olarak ortaya çıkmaktadır: **i.** belirsizlikler; **ii.** çok sözcüklü birimlerin parçası olmak. Türkçede otomatik biçimbilimsel çözümleme ve işaretlemede daha önceki çalışmalarda anılan sorunlar sıklık listelerinde yeralan sözcükbirimlerin çözümlenmesinde de kendisini göstermiştir. Çok daha uzun ve kendi başına bir tartışma gerektiren bu konu bir başka çalışmanın konusudur. Örneğin, eşyazımlı sözcükler için yaklaşık 7,5 milyon sözcüğün TUD içindeki bağlam kullanımına bakarak belirginleştirme yapılması gerektiği için bu tür eşyazımlı sözcüklerin derlem içi belirginleştirme çalışması bir başka proje çalışmanın konusu olarak görülmüştür. Farklı diller için hazırlanan sıklık sözlüklerinde de benzer durumlar tanımlanmış ve ayrıntılı olarak tartışılmıştır (bkz. örn. Leech, Rayson ve Wilson, 2001; *Čermák* ve Křen , 2005).

Burada çalışma sırasında gözlemlenen kimi belirsizliklerin örneklerini anmakla yetineceğiz. Aşağıda yalnızca bir bölümü örneklenen bu belirsizlikler (Aksan, Mersinli, Yaldır ve Demirhan, 2011) Türkçenin düzenli ancak çok zengin biçimbirim kullanım ortamlarından kaynaklanmaktadır.

Tablo 4. Biçimbilimsel Belirsizlik Türleri

Belirsizlik Türü	Sözcük Yapısı	1. Okuma	2. Okuma
Başsözcük	yaz	eylem (yazmak)	ad (mevsim)
Ekler	okulu	3.kişi tekil iyelik eki (onun okulu)	belirtme durumu (Ben okulu seviyorum)
Başsözcük+Ek birleşmesi	alan / al-an	ad (düz, açık yer, saha)	eylem+sıfatlaştırıcı (Kitabı alan çocuk)
Ek+Ek birleşmesi	bil-meden /bil-me-den	bilmeksizin	bilmekten kaynaklı

Belirsizliklerin yanı sıra, sıklık sözlüğü girdisi belirlemede en sık karşılaşılan bir diğer sorun ise çok sözcüklü birimleri ilgilendirmektedir. Az sayıda da olsa kimi sık kullanılan sözcükbirimler gerçekte sıklıklarını bir başka özelliklerinden, çok sözlüklü bir birimin parçası olmaktan almaktadırlar. Tablo 5'de TUD'un yazılı bölümünde en sık karşımıza çıkan üçlü birimlerin kurucu ögelerinde *bir* sözcüğünün kullanım sıklığına bakarsak, başsözcük sıkılığında *bir* sözcüğünün neden 1. sırada karşımıza çıktığını daha rahat anlayabiliriz.

Tablo 5. TUD Yazılı Bölüm: En Sık 5 Üçlü Birim

Sıra	Üçlü Birim	Sıklık
1	**bir** süre sonra	4443
2	**bir** kez daha	4009
3	ne var ki	3363
4	başka **bir** şey	3245
5	ne yazık ki	3025

Çok sözcüklü birim genel olarak, sözcük sınırlarını aşan, sözlüksel, istatistik, sözdizimsel, anlamsal ya da kullanımbilimsel, kendine özgülük taşıyan yapıları (Sag vd., 2002:2) anlatmaktadır. Örneğin, *aynı zamanda, başka bir ifade ile, bir süre sonra*. İşlev sözcülerin, kimi adların, belirteçlerin, ilgeçlerin üst sıralardaki yerini bu sözcüklerin çok sözcüklü birimleri oluşturulmasındaki sık kullanımıyla açıklayabiliriz (bkz. Aksan ve Aksan, 2015). Bu gözlem Türkçe için genel ve güncel sözlükler, özel ve tarihsel sözlüklerin yanı sıra bir ÇSB sözlüğü hazırlamak gerektiğini göstermektedir. *Türkçe'de Çok Sözcüklü Birimler: Derlem-Temelli Yöntem ile Çıkarılmaları, Sayısal Dağılımlarının Hesaplanması, Yapı ve İşlev Özelliklerinin Saptanması, Sözlüğünün Oluşturulması* (Proje no: 115K135, TÜBİTAK-1001) başlıklı projemiz Türkçe doğal dil kullanım ortamlarında dil kullanıcılarının sıklıkla bir araya getirerek oluşturdukları, anlatım gereksinimlerini en etkin biçimde karşılayan, birlikte kullanımları rastlantısal olmayan, birden çok sözcükten oluşan dil birimlerinin neler olduğunun saptanmasını amaçlamaktadır. Bunun için proje;

1. Saptanan çok sözcüklü birimlerin (ÇSB) temel nicelik değerlerinin ve dil kullanım ortamlarındaki sayısal dağılımlarını belirleyecek ve dökümünü çıkartacak;
2. Dökümü çıkartılan bu birimlerin yapı, anlam ve işlev değerleri tanımlanacak, ayrıntılı çözümlemeleri yapılarak ve sınıflamalarını sunacaktır.

6. Sonuç ve Öneriler

Daha önce, bu çalışmanın kullandığı kapsamda ve büyüklükte, temsil yeteneği olan bir dil derlemi üzerinden sıklık çalışması yapılmadığı için, *Türkçe Ulusal Derlemi*'nin 50 milyon sözcüklük veri büyüklüğünün sağladığı, söz-

cükbirim/teksözcük çeşitliliği, başsözcük sayısı, sözcük türüne özgü sıklık dağılımları, çekim eklerinin dizilim ve sıklık özellikleri ilk kez incelenmiş ve bu alanda önemli bir birikim Türkçenin öğretimine katkı sağlayacak bir sözlüğün hazırlanması ve yayınlanması için kullanılmıştır. Sözkonusu sözlüğün eğitimsel değerini şöyle özetleyebiliriz.

• Dil eğitiminde sözcük ve ek sıklıkları bilgileri, ders izlencelerinin oluşturulmasında, okuma öğretimi kitapları geliştirmede, ikinci dil olarak Türkçe öğretim kitaplarında, sözlüklerde ve ikinci dil olarak Türkçe öğrenenlere yönelik testlerin geliştirilmesinde hangi sözcüklerin, sözcük türlerinin ve eklerin seçileceğine karar verilmesinde önemli bir rol oynayacaktır.

• Türkçenin nicelik özellikleri ve özelde sözcük yoğunluğuyla metin türlerinin doğru tanımlanması ve türe özgü sözcüksel sıklık ve dağılımların saptanması okuma eğitimini-öğretimini ve sözcük eğitimi-öğretimini çok daha etkin ve verimli yapacaktır. Okunacak metinlerin seçiminde sahip olması gereken sözcüklerin belirlenmesinde sözcük ve ek sıklığı sözlüğü temel kaynak olacaktır.

• Sıklık Sözlüğü, dil eğitiminde içeriğin örgütlenmesi konusunda da yararlı olacaktır ve ders kitabı ve benzeri dil öğretimi içeriğinin yansız, nesnel ve bilimsel ölçütlerle oluşturulmasıyla daha etkin bir dil öğretimine olanak tanıyacaktır.

Bundan sonraki çalışmalar için aşağıdaki önerileri sunabiliriz:

1. Dilbilimciler ve yazılım mühendisleri arasında daha etkin bir işbirliğinin sağlanması.
2. Belirginleştirme kurallarının arttırmak, daha ayrıntılı tanımlamak ve DDİ araçlarının bir parçası haline getirmek.
3. Türkçenin eşdizimli ögelerinin sunumu içeren sözlüklerin hazırlanmasını sağlamak.

Kaynakça

Aksan, M., Aksan, Y. (2015). Multi-word Units in Genre Specification. *Dil ve Edebiyat Dergisi*, 12(1), 1-42.

Aksan, Y., Yaldır, Y. (2011). Türkçe sözvarlığının nicel betimlemesi., Ç. Sağın-Şimşek ve Ç, Hatipoğlu. (Editörler). *24. Ulusal dilbilim kurultayı bildiri kitabı*. Ankara. ODTÜ Basım İşliği, ss. 377-387.

Aksan, Y., Mersinli, Ü., Yaldır, Y. (2011). İlköğretim Türkçe Ders Kitapları Derlemi ve Türkçe Ulusal Dil Derlemi örneklemindeki sözcük sıklıkları., V.D. Günay, F. Özden, B. Çetin, F. Yıldız (Editörler). *Türkçe* öğretimi üzerine çalışmalar. İzmir. Dokuz Eylül Üniversitesi Yayınları, ss. 397-408.

Aksan, Y. Mersinli, Ü. Yaldır, Y., Demirhan, U. U. (2011). Türkçe Ulusal Dil Derlemi projesi biçimbirim çalışmalarında belirsizliklerin sınıflandırılması ve dağılımı., H. Çubukçu, F. Türkay, E. Örsdemir, D. Sucak (Editörler). *25. ulusal dilbilim kurultayı bildiri kitabı*. Adana. Çukurova Üniversitesi, ss. 219-226

Aksan, Y., Aksan, M., Özel, S.A. vd. (2015). Web tabanlı Türkçe Ulusal Derlemi., *14. Akademik bilişim bildirileri*. ss. 723-730.

Aksan, Y., Özel, S.A., Bektaş, Y. vd. (2015). Türkçe tümcelerin sonunu belirlemede açık kaynak/ ücretsiz yazılımlar ve performans analizleri., *14. Akademik Bilişim Bildirileri*. ss.731-738.

Čermák, F., Křen, M. (2005). New Generation Corpus-based Frequency Dictionaries. *International Journal of Corpus Linguistics*, 10, 453-467.

Čermák, F., Křen, M. (2011). A Frequency Dictionary of Czech. London: Routledge.

Davies, M., Gardner, D. (2010). *A Frequency Dictionary of Contemporary American English*. London: Routledge.

Göz, İ. (2003). *Yazılı Türkçenin Kelime Sıklığı Sözlüğü*. Ankara: Türk Dil Kurumu.

Güngör, T. (2003). Lexical and morphological statistics for Turkish., 18

Mart University (Editör). *International twelfth Turkish symposium on artificial intelligence and neural networks* (TAINN 2003). Çankkale. 18 Mart University.

Gries, S. Th. 2008. Dispersion and Adjusted Frequencies in Corpora. *International Journal of Corpus Linguistics*, 13, 403-437.

Kumova, S., Karaoğlan, B., Dinçer, B.T. (2006). Kelime sayısı-kelime dağarcığı ilişkisinin belirlenmesi., İstanbul University, Natural Language Processing Laboratory (Editör). *Turkish symposium on artificial intelligence and neural networks* (TAINN 2006). İstanbul. İstanbul University.

Leech, G., Rayson, P., Wilson, A. (2001). *Word Frequencies of Written and Spoken English: Based on the British National Corpus*. London: Longman.

Mersinli, Ü., Aksan, M. (2011). Türkçenin biçimbirim ve sözcük türü işaretlemesi., Ç. Sağın-Şimşek ve Ç, Hatipoğlu. (Editörler). *24. Ulusal dilbilim kurultayı bildiri kitabı*. Ankara. ODTÜ Basım İşliği, ss. 213-218.

Ölker, G. (2011). *Yazılı Türkçenin Kelime Sıklığı Sözlüğü (1945-1950 Arası)*. Konya: Kömen Yayınları.

Pierce, J. E. (1961). A Frequency Count of Turkish Affixes. *Anthropological Linguistics*, 3, 31-42.

Pierce, J. E. (1962). Frequencies of Occurrence for Affixes in Written Turkish. *Anthropological Linguistics*, 4, 30-41.

Sag, I. A., Bond, F., Copestake, A. ve Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. *Lecture Notes in Computer Science*, 2276, 1-15.