Türkiye Türkçesinde Eşdizimlerin İstatistiksel Yöntemlerle Belirlenmesi

Senem Kumova Metin*

Bahar Karaoğlan**

Özet

Eşdizim, sözcüklerin bir anlam bütünlüğü oluşturmak üzere şans eseri olmayacak sıklıkla bir araya geldiği sözcük birliğidir. Eşdizimlerin belirlenmesi, Türkçe metinlerin otomatik olarak işlenmesi ve çevirilerinin yapılması, Türkçe dilinin eğitimi gibi çeşitli alanlardaki faydaları sebebiyle Türkçe doğal dil işleme çalışmalarında önemli bir konudur. Bu çalışmada, Türkçe bir derlemde (*corpus*) eşdizimlerin otomatik olarak belirlenmesi için çeşitli istatistiksel teknikler, gözlenme sıklığı (*occurrence frequency*), noktasal karşılıklı bilgi katsayısı (*pointwise mutual information*) ve hipotez testleri uygulanmıştır. Eşdizimlerin belirlenmesinde gövdelemenin etkisinin araştırılması amacıyla sözcüklerin yanısıra bu sözcüklere ait sözcük gövdeleri üzerinde de çalışılmış, yöntemlerin başarımı F-ölçütü (*F-measure*) ile değerlendirilmiştir. Ki-kare hipotez testi ve noktasal karşılıklı bilgi katsayısı Türkiye Türkçesinde eşdizimlerin belirlenmesi konusunda diğer yöntemlere göre daha başarılı olmuştur. Ayrıca, gövdelenmiş sözcüklerden oluşan veri kümelerinde başarılı ve başarısız olarak kabul edebileceğimiz yöntemler arasındaki farkın daha net ortaya çıktığı görülmüştür.

Anahtar Kelimeler

Eşdizim, Türkiye Türkçesi, Doğal Dil İşleme, Derlem

* Yrd. Doç. Dr., İzmir Ekonomi Üniversitesi, Mühendislik ve Bilgisayar Bilimleri Fakültesi, Yazılım Mühendisliği Bölümü – İzmir/Türkiye
senem.kumova@ieu.edu.tr

** Prof. Dr., Ege Üniversitesi, Uluslararası Bilgisayar Enstitüsü – İzmir /Türkiye
bahar.karaoglan@ege.edu.tr

1. GİRİŞ

Eşdizimler sözcüklerin bir araya gelişlerinin şansa bağlanmayacak sıklıkta görüldüğü geleneksel sözcük birlikleridir. Eşdizim kavramı ilk kez İngiliz dilbilimci Firth (1951) tarafından yayınlanan "Modes of Meaning" adlı kitapta yer almıştır. Firth (1951) bir sözcüğün ancak, kendisine eşlik eden sözcük ile değerlendirilebileceğini vurgulamış ve "Verilen bir sözcüğün eşdizimleri bu sözcüğün alışılagelmiş veya geleneksel pozisyonlarıdır" şeklinde bir tanım ortaya koymuştur. Daha sonra Sinclair (1991) eşdizimi, bir metin içerisinde iki veya daha çok sözcüğün kısa mesafede yer alması şeklinde tanımlamıştır. Hoey (1991) ise istatistiksel bir yaklaşımla eşdizime "Kendi bağlamında raslantısal olamayacak miktarda bir arada görülen sözcükler arası ilişkiye verilen isim" demiştir. Literatürde eşdizim kavramının farklı bakış açılarıyla ele alındığı (Özkan 2007) ve farklı şekillerde tanımlandığı birçok çalışma yer almaktadır. Örneğin, eşdizimlilik kavramı için Aksan (2011) "Belli bağlamlarda, bir sözcüğün belirli bir aralık içinde en sık birlikte kullanıldığı sözcük" ifadesini kullanmıştır. Özkan (2010) ise Baker vd. (2006) çalışmasında yer alan "Sözcükbirimlerin anlamsal ya da dilbilimsel birlikteliklerinden kaynaklanan ve kullanım sıklığına bağlı olarak sözlükbirimsel özellik taşıyabilen söz birlikleri (Sterkenburg 2003)" ve "Sözdizimsel olarak anlamsal sözcük birliktelikleri (Hartmann 1998, Tony 2006)" şeklindeki iki tanımı vermektedir.

Eşdizimler nedensiz ve çoğunlukla anlamsal olarak açık olmasalar da yazım ve konuşmada önemli bir anlamsal etkiye sahiptirler. Bu sebeple eşdizimlerin belirlenmesi pek çok doğal dil işleme çalışmasında; doğal dil üretme, makineli çeviri (*machine translation*), anlam belirsizliğinin giderilmesi, sözcük türü bulma ve bilgi çıkarsama; önemli bir konudur. Bu denli geniş bir uygulama alanına hizmet eden eşdizimlerin belirlenmesi amacıyla birçok istatistiksel ve kural tabanlı yöntem geliştirilmiştir. Kural tabanlı yöntemler sözcük türü işaretlenmiş veriler üzerinde çalışan ve metin üzerinde bir grup öncül işlemin gerçeklenmesini gerektiren yöntemlerdir. İstatistiksel yöntemler ise bir çeşit gözlenme sıklığı bilgisini temel alarak verilen bir derlemde eşdizimleri belirleyen yöntemlerdir. En sıklıkla kullanılan istatistiksel teknikler gözlenme sıklığı, noktasal karşılıklı bilgi katsayısı (Church ve Hanks 1990) ve hipotez testleridir (log-olabilirlik (*log-likelihood*), ki-kare, t-testi, vd.). Smadja'nın Xtract'ı (1993), Kita vd. yöntemi (1994), Shimo-

hata vd. teknikleri (1997) de literatürde araştırmacıların adlarıyla anılan önemli çalışmalardandır.

Bu çalışmada, Türkiye Türkçesiyle yazılmış metinler içeren bir derlemde eşdizimlerin otomatik olarak belirlenmesi için bir takım istatistiksel teknikler uygulanmış ve tekniklerin başarımları bilgi geri getirim alanında sıklıkla kullanılan F-ölçütü ile değerlendirilmiştir. Derlemde yer alan sözcükler ve bu sözcüklerin gövdelerinde ilgili yöntemler uygulanarak gövdelemenin eşdizim belirlenmesi üzerindeki etkisi araştırılmıştır. Bu çalışma sonuçlarının, Türkiye Türkçesi başta olmak üzere eklemeli diller alanında yürütülen doğal dil işleme araştırmalarına katkıda bulunacağı umulmaktadır.

2. KURAMSAL ÇERÇEVE

2.1. Tanımlar - Eşdizim: Eşdizimlilik literatüründe yer alan pek çok farklı tanımdan da anlaşılacağı üzere aslında eşdizim oluşturmak için kesin kurallar yoktur ve dil kendi gelişimi içinde eşdizimleri oluşturur[1]. Her ne kadar araştırmacıların ortak olarak kabul ettikleri bir eşdizim tanımı olmasa da eşdizim özellikleri için farklı çalışmalarda ortak olarak rastlanan özellikler şöyledir:

Eşdizimler sıklıkla yinelenir: Yinelenme özelliği, eşdizimlerin diğer sözcük birliklerinden ayırt edilmelerini sağlayan ve ölçümü en kolay olan özellikleridir. Bu sebeple, eşdizim belirleme tekniklerinin neredeyse tümünde eşdizimin rastlanma sıklığıyla ilgili bir bilgi yer almaktadır (Bisht vd. 2006, Smadja 1993, Church ve Hanks 1990, Hindle 1990, Dunning 1993).

*Eşdizimler nedensiz ve dil bağımlıdırlar***:** Dil içinde hangi sözcüklerin eşdizim oluşturacağı hangilerinin oluşturmayacağı; bir sözcüğün milyonlarca sözcük içinden hangisini seçip eşdizim oluşturacağı konusunda bilinen bir kural yoktur. Örneğin "kör talih" Türkçe'de sıklıkla kullandığımız eşdizimlerdendir. Ancak bu eşdizimde "kör" kelimesi yerine eşanlamlı olmasına rağmen niçin "âmâ" sözcüğünü kullanmadığımız açıklanamamaktadır. Ayrıca milletlerin kültürel ve sosyal gelişimleri eşdizimlerin farklılık göstermesine sebep olmaktadır. "kör talih" eşdizimi İngilizce'de "kötü şans (*bad luck*)" olarak geçmektedir.

Eşdizimler birim bloklar oluştururlar: Anlam bütünlüğünü dikkate alan doğal dil işleme çalışmalarında birim bloklar anlam bütünlüğü bulunan bir sözcük veya sözcük grubu olarak tanımlanır. Bu sözcük veya sözcük grubu,

cümle veya cümle öğesi olarak görev yapar. Özellikle anlam belirsizliğinin giderilmesi, cümle öğelerinin saptanması, makine çevirisi gibi anlam bütünlüğünün önemsendiği çalışmalarda birim bloklar önemlidir. Örneğin İngilizcedeki "lady killer" eşdizimi bir bütün (blok) olarak kabul edilmeyerek Türkçeye "kadın katili" şeklinde çevrilebilir. Oysa bu eşdizim bir birim olarak kabul edildiğinde "çapkın" kelimesine denk gelir.

Eşdizimler alan bağımlıdırlar: Eşdizimler dil bağımlı oldukları gibi aynı zamanda spor, sanat, kültür, bilim vb. gibi alanlar içinde de özelleşirler. Smadja (1993) çalışmasında denizcilik alanını örnek göstermiştir. "ıslak giysi" ve "kuru giysi" eşdizimleri denizcilik alanında gerçekten ıslak veya kuru olan giysileri ifade etmez. Bu giysiler deniz suyunun vücuda temasını belirli ölçülerde engelleyen özel giysilerdir.

Eşdizim tanımı halen tartışılmakta olan bir konu olması sebebiyle, bu çalışma kapsamında eşdizim olarak kabul edilen sözcük birlikleri şöyle tanımlanmıştır:

- Deyimler ve bileşik fiiller (örneğin *günah çıkarmak, karar vermek*)
- Sıklıkla kullanılan tamlamalar, alan bağımlı terimler (örneğin *beyaz peynir, cinayet zanlısı*)
- Sıklıkla kullanılan söz öbekleri ve bağlaçlar (örneğin *her şey, ya da*)
- Adlandırılmış varlıklar, makam-pozisyon vb. adları, kısaltmalar (örneğin *Beyaz Saray, genel müdür, prof. dr.*)

2.2. Türkiye Türkçesinde Eşdizim Belirleme Çalışmaları: Türkçe, dil yapısında yer alan pek çok yapım ve çekim eki sebebiyle morfolojik olarak çok üretken bir dildir. Türkçe içinde herhangi bir kök veya gövdeden teorik olarak milyonlarca farklı sözcük türetmek mümkündür. Bu üretkenlik, özellikle hesaplamalı dil bilim (*computational linguistics*) alanında yapılan çalışmaların zaman ve uzay karmaşıklığını yükseltmektedir. Uygulamalardaki karmaşıklığın yanı sıra bir yöntem veya model geliştirilirken dikkate alınan birim (örneğin sözcük, sözcük gövdesi) bir başka uygulamada değiştirilirse ilgili yöntemin veya modelin farklı sonuçlar üretmesi de mümkündür.

Türkçe üzerine yapılan çalışmaların bir kısmı eşdizim kavramının çeviri veya yabancı dil öğretiminde önemini tartışan veya belirli bir sözcüğün eşdizimlilik özelliğini araştıran çalışmalardır (Özkan 2007, Sarıkaş 2006, Taşığüzel 1988).

Eşdizim kavramını genişleterek İlköğretim Türkçe Ders Kitapları'ndaki çok sözcüklü kullanımları belirlemeye yönelik bir araştırma Mersinli ve Demirhan (2010) çalışmasında bulunmaktadır. Bu araştırmanın ders kitabı hazırlama ve değerlendirme süreçlerine katkı sağlaması amaçlanmıştır.

Eşdizimlilik kavramının ayrıntılı bir şekilde incelendiği Özkan (2010) çalışmasında ise Türkiye Türkçesinde sıfatların eşdizim sözlüğünün oluşturulması yöntem ve uygulama açısından değerlendirilmiştir.

Doğal dil işleme yöntemleri ve bilişim teknolojilerinden faydalanılarak Türkiye Türkçesinin Eşdizim Sözlüğü'nün oluşturulmasına yönelik bir diğer çalışma Özkan (2012) tarafından sunulmuştur. Özkan (2012) ilgili sözlüğün oluşturulmasındaki temel amaçlarını "Güncel Türkçe Sözlük'ün fiil, zarf, sıfat, isim temelli "derlem-denetimini" yapmak, ana dili ve ikinci dil öğretiminde önemli bir yere sahip olan eşdizimsel yapıları Türkçe için bu eksende tespit etmek" şeklinde vermiştir.

Doğal dil işleme yöntemlerinin kullanıldığı bir diğer çalışmada, Oflazer vd. (2004) Türkçe'de çoklu sözcük birimlerinin (*multi-worded units*) belirlenmesi için kural tabanlı bir çözüm önermişlerdir. Bu sistem sözcük türleri işaretlenmiş ve sözcük ekleri belirlenmiş bir derlem üzerinde çalışmaktadır. Oflazer vd. (2004) çoklu sözcük birimlerini 4 farklı grupta değerlendirmiştir: sözcüklerin ek almadığı sabit ifadeler, sözcüklerin bir kısmının ek almadığı bir kısmının alabildiği yarı sabit ifadeler, sözcük tekrarları veya zıtlıkları içeren ifadeler, adlandırılmış varlıklar. Çalışmada belirli morfolojik desenlerin 1100 kural çerçevesinde değerlendirilmesi sonucunda çoklu sözcük birimlerinin metin içerisinde belirlenmesi amaçlanmıştır.

3. YÖNTEM

Eşdizimlerin belirlenmesinde kural tabanlı ve istatistiksel yöntemler mevcuttur. Bu çalışmada istatistiksel yöntemler ele alınmıştır. İzleyen alt bölümlerde çalışma içinde uygulanan istatistiksel teknikler; gözlenme sıklığı, noktasal karşılıklı bilgi katsayısı, hipotez testleri, ortalama-varyans yöntemi, Smadja yöntemi ve eşdizim eğilimi yöntemi (Kumova Metin vd. 2011) tanıtılmaktadır.

3.1. Gözlenme Sıklığı: İki veya daha fazla sözcüğün birlikte gözlenme sayısını temel alan yöntem uygulanırken derlemde yan yana bulunan (*n-gram*) veya bir pencere dâhilinde bir arada bulunan sözcüklerin gözlenme sıklık-

ları ölçülür. Derlemdeki bu sözcük birlikleri sıklık değerleri azalacak şekilde listelenir, bu liste eşdizim adaylarını içermektedir. Listede yüksek sıklık değerine sahip olan adayların eşdizim olduğu kabul edilir. Bu yöntemdeki en büyük problem gerçek eşdizimlilikleri, diğer sözcük birliklerinden ayıran eşik değerin belirlenmesi aşamasıdır. Yöntemin dezavantajı ise çok sıklıkla gözlenen eşdizim adayları içinde işlev kelimelere (örneğin *bir, bu, ve, şey, gibi*) rastlanmasıdır (Manning ve Schütze 1999). Bu sebeple sıklık değerine bağlı olarak hazırlanan listeler sözcük türü filtresi (Justeson ve Katz 1995) gibi çeşitli filtrelerden geçirilir. Örneğin, bu tip bir filtre sayesinde sadece isim tamlamalarının eşdizim adayı olarak değerlendirilmesi sağlanabilir.

3.2. Noktasal Karşılıklı Bilgi Katsayısı: Karşılıklı bilgi katsayısı, enformasyon teorisindeki tanımıyla, iki rassal değişkenin noktasal karşılıklı bağımlıklarını gösteren bir değerdir. İki değişkenin bir arada görülme olasılığının ayrı ayrı görülme olasılıklarına bölünmesiyle elde edilen değerin iki tabanında logaritması alınarak bu bağımlılığın bit cinsinden değeri ifade edilir.

Hesaplamalı dil bilim çalışmalarında, eşdizim oluşturan sözcüklerin bir arada bulunuşlarının tesadüfî olmadığı fikrinden yola çıkarak sözcükler arası bağımlılıkların miktarını saptamak için noktasal karşılıklı bilgi katsayısı kullanılmaktadır (Church ve Hanks 1990, Hindle 1990). Bu yöntemde bir arada görünen iki sözcük için her sözcüğün kendi başına görülme olasılığı ile birlikte görülme olasılığı arasındaki ilişki hesaplanır. Bu ilişkiye dayanarak eşdizimlilik kararı verilir. Denklem 1'de w_1 ve w_2 sözcükleri için noktasal karşılıklı bilgi katsayısı, $I(w_1w_2)$, verilmiştir.

$$I(w_1w_2) = \log_2 \frac{P(w_1w_2)}{P(w_1) * P(w_2)} \qquad (1)$$

Toplam sözcük sayısı N olan bir derlemde $f(w_1)$, $f(w_2)$ ve $f(w_1w_2)$ sırasıyla w_1, w_2 sözcükleri ve w_1w_2 ikilisinin sıklık değerini gösterirken, her bir sözcüğün ve ilgili ikilinin gözlenme olasılıkları şu şekilde tanımlanır:

$$P(w_1) = \frac{f(w_1)}{N}$$
$$P(w_2) = \frac{f(w_2)}{N}$$
$$P(w_1 w_2) = \frac{f(w_1 w_2)}{N}$$

Denklem 1' de w_1 ve w_2 sözcükleri birbirinden bağımsız ise birlikte gözlenme olasılıkları, sözcüklerin ayrı ayrı gözlenme olasılıklarının çarpımına eşittir ($P(w_1 w_2) = P(w_1) \cdot P(w_2)$). Dolayısıyla eşdizim oluşturmayan w_1 ve w_2 sözcükleri için noktasal karşılıklı bilgi katsayısı sıfır olmaktadır ($I(w_1 w_2) = 0$). Sözcüklerin bağımlı olduğu yani eşdizim oluşturduğu durumda ise noktasal karşılıklı bilgi katsayısı sıfırdan uzaklaşacaktır. Bu sebeple herhangi bir sözcük birliği için noktasal karşılıklı bilgi katsayısı sıfır değerinden ne kadar uzaklaşır ise bu birliklerin eşdizim özelliğinin o denli arttığı kabul edilir.

Yöntemin uygulanmasında gözlenme sıklığı yöntemine benzer şekilde tüm derlemde sözcük birlikleri için noktasal karşılıklı bilgi katsayısı ölçülür, bu değer azalacak şekilde sözcük birlikleri listelenir. Listede yüksek karşılıklı bilgi katsayısına sahip olan birlikler eşdizim olarak kabul edilir.

3.3. Hipotez testleri: Bir sözcük birliğinin eşdizim olup olmadığının belirlenmesi için sözcüklerin bir arada bulunuşlarının şans eseri olmadığının ispatı gereklidir. Sözcükler arası bağımlılığı göstermek için çoğunlukla sözcükler arası bağımsızlık test edilir. Eşdizimliliğin belirlenmesinde kullanılan hipotez testleri takip eden bölümlerde kısaca anlatılmaktadır.

3.3.1. t-testi: t-testinde sıfır hipotezi örneğin μ ortalamaya sahip normal dağılımdan çekildiğini varsayar. Bu sebeple, gözlenen ortalama değer beklenen ortalama değerden (μ) farklılık gösterir ise sıfır hipotezi reddedilir. Test beklenen ve gözlenen ortalama değerlerin farkını, örneğin varyansına göre ölçeklendirerek değerlendirir. t değeri şu şekilde hesaplanır:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \qquad (2)$$

Denklem 2'de $\bar{x}$ örnek ortalama değeri (gözlenen ortalama değer), s^2 varyans, N örnek büyüklüğü ve μ ise beklenen ortalama değerdir.

Eşdizimliliğin belirlenmesi için t-testinin kullanılması durumunda sıfır hipotezi sözcüklerin bağımsız olduğunu ifade eder. $f(w_1)$, $f(w_2)$ ve $f(w_1w_2)$ sırasıyla w_1, w_2 sözcükleri ve w_1w_2 ikilisinin sıklık değerini, N derlemdeki toplam sözcük sayısını gösterirken, her bir sözcüğün ve ilgili ikilinin gözlenme olasılıkları şu şekilde tanımlanır:

$$P(w_1) = \frac{f(w_1)}{N}$$
$$P(w_2) = \frac{f(w_2)}{N}$$
$$P(w_1w_2) = \frac{f(w_1w_2)}{N}$$

Bu durumda w_1w_2 ikilisinin t-testi ile eşdizimlilik sınamasında sıfır hipotezi $H_0 : P(w_1w_2) = P(w_1) \cdot P(w_2)$ olur. Eğer sıfır hipotezi doğru ise rastgele sözcük ikilileri içinde w_1w_2 ikilisinin seçilmesi ve bu ikilinin başarılı sonuç, diğer tüm ikililerinin başarısız sonuç olarak kabul edildiği bir Bernoulli deneyi söz konusudur. Bu durumda dağılımın ortalama değeri (t-testi için beklenen değer) $\mu = p = P(w_1)P(w_2)$, dağılımın örnek ortalaması (t-testinde gözlenen değer) $\bar{x} = P(w_1w_2)$ ve varyansı $s^2 = p(1-p) = P(w_1w_2)(1-P(w_1w_2)) \approx P(w_1w_2)$ olur (Manning ve Schütze 1999). İlgili ikili için t değeri şu şekilde hesaplanır:

$$t = \frac{P(w_1w_2) - P(w_1) \cdot P(w_2)}{\sqrt{\frac{P(w_1w_2)}{N}}} \qquad (3)$$

Bir sözcük ikilisi için hesaplanan t değeri serbestlik derecesi= $N-1$ için belirlenen güven düzeyindeki (*confidence level*) kritik değerden büyük ise sıfır hipotezi reddedilir. Bu durumda ilgili ikilinin rastgele bir arada bulunmadığı sonucu ortaya çıkar. Derlemde bulunan tüm sözcük ikilileri için t de-

ğeri hesaplanıp, tüm ikililer birbiriyle kıyaslanabilir. Bir sözcük ikilisinin t değerinin yüksek olması eşdizim ihtimalinin aynı derlemdeki diğer ikililere oranla daha yüksek olması anlamına gelir.

3.3.2. Pearson χ^2 (ki-kare) testi: χ^2 (ki-kare) testi sözcük birliklerinin beklenen ve gözlenen sıklık değerlerine bağlı olarak eşdizimliliklerinin değerlendirildiği bir testtir. Bu testte, sözcük birliğinde yer alan sözcüklerin sıklık değerleri Tablo 1'de verildiği üzere 2x2'lik bir tabloya yerleştirilir. Tablo 1 için sıfır hipotezi "beyaz" ve "saray" sözcüklerinin bağımsız olduğudur ve beklenen sıklık değeriyle ifade edilir. "beyaz saray" ikilisi için ölçülen yani gözlenen ortalama değer beklenen değerden farklılaştıkça ikilinin eşdizim olma ihtimali yükselir.

Tablo 1: *"beyaz" ve "saray" Sözcükleri için Gözlenen Sıklık Değerlerini İçeren 2x2'lik Tablo*

	w_1=*beyaz*	$w_1 \neq$ *beyaz*
w_2=*saray*	8 (*beyaz saray*)	*4667 (örneğin, kervan saray)*
$w_2 \neq$ *saray*	*15820 (örneğin, beyaz tül)*	*14287181 (örneğin, kedi tüyü)*

Ki-kare istatistiği, gözlenen (O_{ij}) ve beklenen (E_{ij}) değerler arasındaki farkların tablonun tüm hücreleri için toplamını ifade eder. Hesaplanan değer aynı zamanda beklenen değer ile ölçeklendirilir.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_j)^2}{E_j} \qquad (4)$$

Denklem 4'de i tablodaki satır, j ise sütun indeksini simgeler. Tablodaki her bir hücre için beklenen değer, E_{ij}, ilgili hücrenin satır toplamının sütun toplamı ile çarpılıp tablo toplamına bölünmesiyle hesaplanır. Örneğin "beyaz saray" ikilisi için beklenen değer

$E_{11} = \frac{(8+4667)(8+15820)}{(8+4667+15820+14287181)}$ olur.

Eşdizimlilik kararında her bir sözcük ikilisi için χ^2 değerinin belirlenen güven düzeyindeki kritik değerden büyük olması sınanır (2x2'lik bir tablo için serbestlik derecesi=1'dir). χ^2 değeri kritik değerden büyükse ilgili ikilinin eşdizim olabileceği kabul edilir.

Derlemde bulunan tüm sözcük ikilileri için χ^2 değeri hesaplanarak ikililer χ^2 değerleri azalacak sırada listelenir. χ^2 değeri yüksek olan adaylar eşdizim olmaya en yatkın adaylar olarak kabul edilir.

3.3.3. Log-olabilirlik testi: Log-olabilirlik yöntemi Dunning tarafından önerilen (1993) bir hipotez testidir. Dunning yönteminde eşdizimliliğin belirlenmesi için, w_1w_2 sözcük ikilisinin gözlenme sıklığı için iki alternatif tanım verilir:

Hipotez 1 : $P(w_2 / w_1) = p = P(w_2 / w_1{}^c)$

Hipotez 2 : $P(w_2 / w_1) = p_1 \neq p_2 = P(w_2 / w_1{}^c)$

Bu hipotezlerde, $P(w_2 / w_1)$ terimi w_1 sözcüğünün gözlendiği durumda w_2 sözcüğünün gözlenme olasılığını, $P(w_2 / w_1{}^c)$ terimi w_1 sözcüğünün gözlenmediği, $w_1{}^c$, durumda w_2 sözcüğünün gözlenme olasılığını ifade eder. Sonuçta hipotez 1, w_2 kelimesinin gözlenme durumunun, w_1 sözcüğünden bağımsız olduğunu, hipotez 2 ise w_1 sözcüğünün gözlendiği ve gözlenmediği durumlarda w_2 sözcüğünün gözlenme ihtimalinin farklılaştığını belirtir. Eğer hipotez 1 kabul edilirse, sözcük ikilisi bir eşdizim değildir, eğer hipotez 2 kabul edilirse ikili bir eşdizim oluşturur. N adet sözcük içeren bir derlemde p , p_1 ve p_2 değerleri şöyle hesaplanır:

$$p = \frac{c_2}{N}$$

$$p_1 = \frac{c_{12}}{c_1}$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

c_1, c_2, c_{12} sırasıyla w_1, w_2 sözcükleri ve w_1w_2 ikilisinin derlemdeki gözlenme miktarlarıdır.

N adet sözcük içeren bir derlemde w_1 , w_2 sözcükleri ve w_1w_2 ikilisinin sırasıyla c_1 , c_2 ve c_{12} kere gözlenmesi olaylarını binom dağılım ile

$$(b(k,n,x) = \binom{n}{k} x^k (1-x)^{n-k})$$

ifade edersek hipotez 1 için $L(H_1) = b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)$, hipotez 2 için ise $L(H_2) = b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2)$ olabilirlik değerleri elde edilir[2]. Bu durumda log-olabilirlik oranı, λ , şu şekilde tanımlanır:

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)} = \log \frac{b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2)} \qquad (5)$$

Mood (1974) $-2\log\lambda$ dağılımının asimptotik olarak χ^2 dağılımı olduğunu göstermiştir (Dunning, 1993). Bu sebeple hesaplanan $-2\log\lambda$ değeri, verilen güven düzeyinde χ^2 kritik değerinden (serbestlik derecesi =1) küçük ise bağımsızlığı simgeleyen hipotez; aksi durumda ise w_1w_2 ikilisinin bir eşdizim olduğunu belirten hipotez kabul edilir. Bu sebeple log-olabilirlik oranı bir hipotezin diğerine oranla ne kadar kabul edilebilir olduğunu gösteren bir değerdir. Yöntemin uygulanmasında derlemdeki tüm kelime ikililerinin $-2\log\lambda$ hesaplanır ve ikililer ilgili değer azalacak şekilde listelenir.

3.4. Ortalama–Varyans Yöntemi: Ortalama-varyans yöntemi eşdizimlerin saptanmasında sıklık değeri kadar sözcükler arasındaki mesafenin de önemli olduğu fikrinden ortaya çıkmıştır. Özellikle sıklık yönteminin yakalamakta başarısız olduğu aralarına farklı sözcükler girebilen eşdizimleri saptamakta daha başarılı olduğu hâlihazırda İngilizce üzerine yapılan çalışmalarda görülmüştür (Manning ve Schütze 1999).

Yöntem tüm sözcük ikililerinin ve sözcükler arası uzaklıkların listelenmesini gerektirir. Bir sözcük ikilisi için verilen pencerede farklı uzaklıklarda görülme değerlerinin ortalaması ($\bar{d}$) ve varyansı (s^2) denklem 6 ve 7 de verilen şekilde hesaplanır (Manning ve Schütze 1999) .

$$\bar{d} = \frac{\sum_{i=1}^{i=n} d_i}{n}. \qquad (6)$$

$$s^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1} \quad (7)$$

Denklem 6 ve 7'deki d_i terimi, derlemde n kere birlikte gözlenen bir sözcük ikilisi için derlemde beraber gözlendikleri i'inci yerde birinci sözcük ile ikinci sözcük arasındaki mesafeyi (uzaklığı) sözcük miktarı cinsinden ifade eder. Örneğin w_j ve w_k sözcüklerinin eşdizimliliğinin araştırıldığı durumda $w_1 w_2 w_3 ... w_j w_{j+1} ... w_k w_{k+1} ... w_m$ örnek dizinindeki uzaklık değeri $k - j$ olur.

Örneğin bir derlemde "kafa ye" birliğinin rastlandığı yerler sırasıyla şu şekilde ise

1.İşte kafayı şimdi yiyeceğim

2.Bu problem bana kafayı yedirtti

3.Kafayı yavaş yavaş yedi

4.Kafayı yemeden bir atlatsak çok rahatlayacağız

İlk gözlenme için uzaklık değeri $d_1 = 2$, ikinci ve dördüncü için $d_2 = d_4 = 1$, üçüncü için ise $d_3 = 3$ olmaktadır. Bu durumda ortalama değer

$$\bar{d} = \frac{(2+1+3+1)}{4} = 1.75$$

şeklinde hesaplanır. Bu değer "kafa ye" birliği için ikinci sözcüğün birinci sözcükten ortalama olarak 1.75 sözcük uzaklıkta yer aldığını gösterir. Varyans ise

$$s^2 = \frac{(2 - 1.75)^2 + (1 - 1.75)^2 + (3 - 1.75)^2 + (1 - 1.75)^2}{3}$$

şeklinde hesaplanır. Varyans değeri iki sözcük arasındaki uzaklığın ortalama değerden ne kadar saptığını gösterir. Bu değerin sıfır (sıfıra yakın) olması sözcüklerin sürekli aynı uzaklıkta ($\bar{d}$) yer aldıklarının dolayısıyla eşdizimlilik özelliklerinin bulunduğunun bir göstergesidir.

3.5. Smadja Yöntemi: Smadja'nın (1993) Xtract isimli çalışmasında sözcükler arası eşdizimlilik farklı eşdizimlilik özelliklerinin sınandığı 3 temel aşamayla belirlenir. Bir derlemdeki iki sözcükten oluşan tüm eşdizimleri belirlemek için tüm sözcük çiftleri aynı sınamalara tabi tutulur, sınamalarda belirli eşik değerleri geçen çiftler bir sonraki aşamada değerlendirilirler. Değerlendirme aşamaları sırayla şu şekildedir:

Aşama 1: Bir eşdizimin gözlenme sıklığı (gücü), derlemde eşdizimin ilk sözcüğüyle başlayan tüm sözcük çiftleri dikkate alınarak hesaplanan ortalama sıklık değerinden yüksek olmalıdır (Smadja 1993). Bu özelliğin sınanması için Smadja sözcük çiftinin gücü (*strength of word pair*), k, kavramını ortaya koymuştur. k değeri bir sözcük ikilisinin eşdizimlilik olasılığının aynı ilk sözcüğü içeren diğer ikililer vasıtasıyla hesaplanmış değeridir. Denklem 8 herhangi bir *i* sözcük ikilisi için güç değerini vermektedir. Denklem 8'de f ilgili ikilinin gözlenme sıklık değeri, $\bar{f}$ bu sözcük ikilisinin ilk sözcüğü ile başlayan tüm ikililerin ortalama sıklık değeri, σ ise sıklık standart sapma değeridir (Smadja 1993).

$$k = \frac{f - \bar{f}}{\sigma} \qquad (8)$$

Belirli bir sözcük çifti için güç ölçümünün yapılabilmesi, bir pencere dâhilinde çiftin ilk sözcüğü ve ondan sonra gelen tüm sözcüklerin gözlenme sıklıklarının ölçülmesiyle mümkündür. Tablo 2'de örnek olarak "maliye" sözcüğü verilmiştir (değerler herhangi bir derlemden alınmamıştır). Bu örnek için derlemde 10 sözcüklük bir pencere dâhilinde eşdizim adayı çifte ("maliye bakan") ait ilk sözcüğün ("maliye") bir arada gözlendiği tüm farklı sözcükler belirlenmiş ve oluşan ikililerin sıklık değerleri (f) ölçülmüştür. Örneğin aday "maliye bakan" ikilisi $f =$ 15 kez gözlenmiştir. Tablo 2'de son 10 sütunda yöntemin bir sonraki aşamasında kullanılan mesafe-sıklık değerleri verilmiştir. Örneğin, "maliye bakan" ikilisi 1 birim mesafede yani yan yana $p_1 = 12$, 2 birim mesafede $p_2 = 2$, 3 birim mesafede $p_3 = 1$ kez gözlenmiştir. Bu ikili 4-10 birim mesafe aralığında ise gözlenmemiştir.

Tablo 2: *10 Sözcüklük Bir Pencerede "Maliye"Sözcüğü İle Başlayan İkililere Ait Sıklık Değerleri*

Sözcük1	Sözcük2	İkili gözlenme sıklığı (f)	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}
maliye	bakan	15	12	2	1	0	0	0	0	0	0	0
maliye	türkiye	7	1	2	2	2	0	0	0	0	0	0
maliye	üzeri	4	1	0	3	0	0	0	0	0	0	0
maliye	ziyaret	2	0	0	1	1	0	0	0	0	0	0
maliye	yarat	2	0	0	0	1	1	0	0	0	0	0
maliye	yap	1	1	0	0	0	0	0	0	0	0	0

Tablo 2'de yer alan bilgilere dayanarak "maliye" sözcüğü ile oluşturulacak birlikler icin ortalama sıklık değeri

$$\bar{f} = \frac{(15+7+4+2+2+1)}{6} = 5.17$$

şeklinde hesaplanır. Standart sapma değeri ise

$$\sigma = \sqrt{\frac{(15-5.17)^2+(7-5.17)^2+(4-5.17)^2+(2-5.17)^2+(2-5.17)^2+(1-5.17)^2}{6}} = 4.81$$

olmaktadır. Daha sonra ise denklem 8 kullanılarak her bir eşdizim adayının (ikilinin) k değeri belirlenir (örneğin "maliye bakan" ikilisi için $k = 2.043$). k_0 eşik değerinden büyük k değerine sahip adayların eşdizimli olabileceği kabul edilir ve bir sonraki aşamaya geçilir (Smadja 1993). Smadja (1993) pencere büyüklüğünü $W = 10$ olarak alındığı deneysel bir çalışma sonucunda $k_0 = 1$ sabit değerini elde etmiştir.

Aşama 2: Bir eşdizim içinde yer alan sözcükler derlemde sıklıkla aynı dizilimde yani aynı mesafede (uzaklıkta) gözlenmelidirler (Smadja 1993). Bu doğrultuda bir aday ikili için mesafe-sıklık ilişkisi sınanırken ikiliye ait sözcüklerin bir pencere dâhilinde farklı mesafelerde ne sıklıkta gözlendikleri belirlenir. Bu sıklık değerlerinin ortalaması hesaplanır. Gözlenen sıklık değerlerinin ortalamadan ne kadar farklılaştığı (varyansı), U , hesaplanır. Eğer aday ikili tüm uzaklıklarda eşit miktarda gözleniyorsa bu fark düşük, ilgili ikili tek bir uzaklıkta çok sıklıkla gözlenirken diğer uzaklıklarda gözlenmiyor ise fark yüksek olacaktır. Bu farkın yüksek olması adayın eşdizimliliği-

ni, düşük olması ise sözcüklerin rasgele bir arada bulunduklarını gösterir. Denklem 9'da ilgili hesaplama verilmektedir. Denklemde W sözcük cinsinden pencere büyüklüğü, $\overline{p}$ sözcük ikilisi için ortalama ikili sıklığı (toplam ikili sıklık değerinin pencere büyüklüğüne oranı), p_j ise ikilinin j mesafede birlikte gözlenme sıklığıdır.

$$U = \frac{\sum_{j=1}^{5} (p_j - \overline{p})^2}{W}. \qquad (9)$$

Tablo 2'de yer alan "maliye bakan" ikilisi için ilgili değerler

$$\overline{p} = (12+2+1+0+0+0+0+0+0+0)/10 = 1.5$$

$$U = \frac{(12-1.5)^2+(2-1.5)^2+(1-1.5)^2+(0-1.5)^2+(0-1.5)^2+(0-1.5)^2+(0-1.5)^2+(0-1.5)^2+(0-1.5)^2+(0-1.5)^2}{10} = 12.9$$

şeklinde hesaplanır.

Smadja yönteminde tüm ikililerin U değerleri hesaplanarak önceden deneysel olarak belirlenmiş bir $U_0 = 10$ eşik değeriyle kıyaslanır, $U > U_0$ olan ikililerin hep aynı uzaklıklarda rastlanmaları sebebiyle eşdizimlilik özelliklerinin olabileceği kabul edilir ve diğer aşamaya geçilir.

Aşama 3: Smadja yönteminin ilk iki aşamasında hangi sözcük ikilisinin eşdizim olarak seçileceği saptanırken eşdizim adayı ikiliye ait sözcüklerin hangi mesafede eşdizim oluşturduğunu belirleyen özellik denklem 10 ile ifade edilir. Denklemde, $\overline{p}$ aday ikili için pencere dâhilinde hesaplanan ortalama ikili sıklığı, w_1 ikilinin j uzaklıkta birlikte görülme sıklığı ve N terimi uzaklık katsayısını ifade eder.

$$p_j \geq \overline{p} + (k_1 \times \sqrt{U}) \qquad (10)$$

Uzaklık ölçümünde aday ikilide p_j değeri $\overline{p} + (k_1 \times \sqrt{U})$ değerinden ne denli büyükse ilgili ikilinin j mesafesinde eşdizimliliğinin o denli kuvvetli olduğu kabul edilir. Deneysel bir çalışma sonucunda $k_1 = 1$ olduğu belirlenmiştir (Smadja 1993). Örneğin, ilk iki aşamayı geçen eşdizim adayı "maliye bakan" ikilisi için en sıklıkla gözlendiği ilk iki mesafede, $j = 1$ için

$12 \geq 1.5 + \sqrt{12.9}$, $j = 2$ için $2 \leq 1.5 + \sqrt{12.9}$ olur. Bu kıyaslama sonucu $j = 1$ mesafesinde yani sözcüklerin yan yana gözlendikleri durumda "maliye bakan" ikilisi eşdizim oluştururken aralarına bir sözcük girdiği durumda, $j = 2$, eşdizim oluşturmaz.

3.6. Eşdizim Eğilimi Yöntemi: Kumova-Metin ve Karaoğlan (2011) tarafından geliştirilen eşdizim eğilimi yöntemi, sözcükler arası anlam bütünlüğü fikrine dayanmaktadır. Eşdizimi oluşturan sözcükler arasında anlam bütünlüğü olması sebebiyle sözcüklerin birbirini çağrıştırdığı kabul edilmektedir (Kumova-Metin ve Karaoğlan 2011). Bu sebeple bir sözcük birliğinin eşdizim olup olmadığına karar vermek için sözcükler arası ve birliğin diğer sözcüklerle olan ilişkisi değerlendirilmelidir.

Yöntem ard arda gözlenen iki sözcükten (örneğin $w_i w_j$) oluşan bir eşdizim adayı için iki aşamalı bir sınama gerektirir. İlk aşamada derlemde yer alan w_i sözcüğünün herhangi bir eşdizimin ilk sözcüğü olup olmadığının sınanması için derlemde bu sözcükle başlayan tüm ikililer (w_i ve onu takip eden ilk komşu sözcük) belirlenir. Eğer w_i sözcüğünü takip eden çok fazla sayıda farklı sözcük var ise w_i anlam bütünlüğünü tamamlamıştır ve eşdizim oluşturmaz. Aksi durumda, yani w_i sözcüğünü takip eden az sayıda sözcük olduğu durumda ise w_i'nin anlam bütünlüğünü tamamlamadığı varsayılır. İkinci aşamada ise w_i sözcüğünü takip eden her bir sözcük ile eşdizimliliği değerlendirilir. Eğer w_i ile başlayan herhangi bir ikiliyi, örneğin $w_i w_j$, takip eden çok sayıda farklı sözcük var ise bu ikilinin eşdizim olduğu kabul edilir.

Bu yöntemde bir sözcüğe/sözcük ikilisine ait sıklık bilgisinin, farklı komşu sözcük miktarına oranı sözcüğe/sözcük ikilisine ait "eşdizim eğilimi" olarak adlandırılmaktadır. w_i sözcüğünün derlemde gözlenme sıklığı $f(w_i)$, w_i sözcüğüne ait farklı komşu sözcük miktarı $n(w_i)$, eşdizim adayı $w_i w_j$ ikilisine ait gözlenme sıklığı $f(w_i w_j)$ ve bu adayı takip eden farklı sözcük miktarı $n(w_i w_j)$ kabul edilerek; w_i sözcüğüne ait eşdizim eğilimi $T(w_i)$ ve $w_i w_j$ aday ikilisine ait eşdizim eğilimi $T(w_i w_j)$ şu şekilde hesaplanabilir:

$$T(w_i) = \frac{f(w_i)}{n(w_i)} \quad \text{ve} \quad T(w_i w_j) = \frac{f(w_i w_j)}{n(w_i w_j)}$$

$T(w_i)$ değeri ne denli yüksekse w_i sözcüğünün eşdizim oluşturma olasılığı, o denli yüksektir. $T(w_i w_j)$ değeri ise ne denli düşükse, $w_i w_j$ ikilisi o denli eşdizim olma eğilimdedir. Yöntemin uygulanmasında farklı eşdizim belirleme yöntemlerinin aday olarak gösterdiği sözcük ikililerine ait $T(w_i)$ ve $T(w_i w_j)$ değerleri hesaplanır, deneysel olarak belirlenen T_0 eşik değeriyle kıyaslanır. $T(w_i) > T_0$ ve $T(w_i w_j) < T_0$ olan ikililer eşdizim olarak kabul edilir.

4. ÖRNEKLEM: DENEY DERLEMLERİ

Eşdizim belirleme yöntemlerinin Türkiye Türkçesi üzerinde sınanması için deney derlemleri olarak çeşitli dilbilim ve bilgi çıkarsama çalışmalarında kullanılmış olan Bilkent ve OSTAD derlemleri kullanılmıştır.

Bilkent derlemi Bilkent Üniversitesi'nde yürütülen hesaplamalı dil-bilim çalışmalarının bir ürünüdür (Tür vd. 2003). Derlemin içeriği, hazırlandığı yıllarda çıkan gazete yazıları ve makalelerden oluşmaktadır (Dinçer 2004). Bilkent derleminin morfolojik analizi (sözcüklerin türü, kök, gövde, yapım ve çekim eklerinin belirlenmesi vb. gibi) bir sonlu durum makinesiyle (*finite state machine*) yapılmıştır. Derlem ~ 719665 sözcükten oluşmaktadır. Cümle sonlarının belirlenmesi de ilgili makine ile sağlanmıştır (Tür vd. 2003). Derlem, Dinçer (2004) tarafından tekrar düzenlenmiştir, bir takım hatalar ayıklanmıştır. Bu çalışmada derlemin düzenlenmiş olan bu son hali yer almaktadır.

OSTAD, ODTU-Sabancı derlemi, (Oflazer vd. 2003, Atalay vd. 2003) Türkçe üzerine yapılan birçok doğal dil işleme çalışmasında kullanılmış olan ODTU derleminin bir alt kümesinden oluşturulmuştur (~ 46532 sözcük). ODTU derlemi içinden çekilen bu bölümün morfolojik analizi elle yapılmıştır (Oflazer vd. 2003, Atalay vd. 2003).

5. SONUÇLAR VE TARTIŞMA

Eşdizim belirleme yöntemlerinin Türkçe derlemler kullanılarak sınanması üç aşamada gerçekleşmiştir.

Ön değerlendirme: Literatürde yer alan yöntemler deney derlemleri üzerinde uygulanarak, her bir yöntemin kendi içinde Türkiye Türkçesi üzerinde geçerliliği incelenmiştir.

Yöntemlerin kıyaslanması: Türkçe'nin yapısına ve amaca uygunluğu kabul edilen yöntemler bu aşamada birbirleri ile karşılaştırılmıştır. Bu kıyaslamada gövdelemenin yöntemlerin başarısı üzerindeki etkisi de incelenmektedir.

Eşdizim eğilimi yönteminin uygulanması: Son aşama olarak, belirlenen yöntemler üzerine Kumova Metin vd. (2011) tarafından önerilen eşdizim eğilimi yöntemi uygulanarak yöntemlerin ürettiği sonuçlardaki değişimler/gelişimler değerlendirilmiştir.

İzleyen alt bölümlerde deneysel çalışma aşamaları tanıtılmaktadır.

5.1.Ön değerlendirme: Ortalama-varyans yöntemi OSTAD derleminde Sözcük gövdelerinden örnek ikililer üzerinde pencere büyüklüğü=5 alınarak uygulanmıştır. Bu durumda bir aday ikiliyi oluşturan iki sözcük gövdesi arasında en fazla 4 sözcük gövdesi olabilir. Tablo 3'de "yüz yüz", "fark et" ve "neden ol" ikilileri için sıklık ve uzaklık bilgileri verilmiştir. Uzaklık sütunlarında eşdizim adayına ait ikinci sözcük gövdesinin ilk sözcük gövdesinden hangi uzaklıkda kaç defa gözlendiği tutulmaktadır. Örneğin "neden ol" ikilisi için "ol" gövdesi "neden" gövdesinden 7 kez 1 uzaklıkta (yan yana), 1 kez de 5 uzaklıkta görülmüştür.

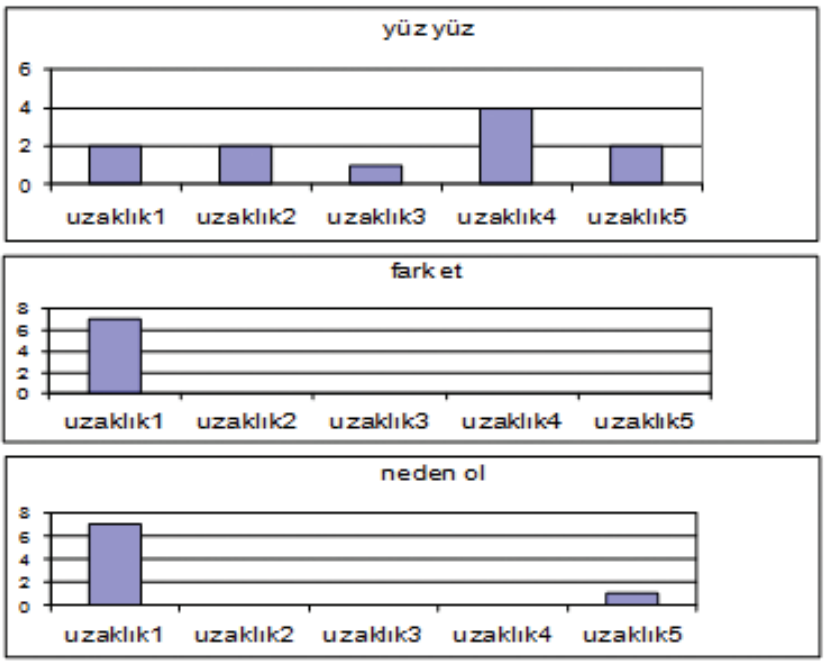
Tablo 3: *OSTAD Derleminden Seçilen Örnek Eşdizim Adayları İçin Bir Arada Gözlenme Sıklık ve Uzaklık Değerleri*

Gözlenme Sıklığı	Sözcük 1	Sözcük 2	uzaklık1	uzaklık2	uzaklık3	uzaklık4	uzaklık5
11	yüz	yüz	2	2	1	4	2
7	fark	et	7	0	0	0	0
8	neden	ol	7	0	0	0	1

Şekil 1'de ise örnek eşdizim adaylarının uzaklıklarına ait histogramlar verilmiştir. Bu histogramlarda belli uzaklık değerlerinde yığılma görülen adaylar eşdizim olarak kabul edilir. Eğer iki sözcük/sözcük gövdesinin tüm uzaklık

değerlerinde görülme olasılığı eşit ise bu grubun eşdizim olmadığı kararına varılır.

OSTAD ve Bilkent derlemlerinde sözcük gövdeleri üzerinde ortalama-varyans yöntemi uygulanarak Tablo 4 ve 5'de verilen sonuçlar elde edilmiştir. Bu tablolarda yöntemin eşdizim adayı olarak belirlediği ikililer gözlenme sıklığına göre azalan sırada listelenmiştir, listelerde ilk 60 ikili yer almaktadır. Listelerde standart sapma (f) ve ortalama değerleri verilmiştir ($\overline{d}$). Bu yöntemde, $\overline{d} = 1$ ve $\sqrt{s^2} \approx 0$ olması ilgili ikilinin yan yana gözlemlenen sözcüklerden/sözcük gövdelerinden oluşan bir eşdizim (örneğin "karar ver"), $\overline{d} \neq 1$ ve $\sqrt{s^2} \approx 0$ olması ise aralarına farklı sözcükler giren sözcüklerden/sözcük gövdelerinden oluşan bir eşdizim (örneğin "üst koy") olduğunu göstermektedir.



Şekil 1. OSTAD derlemi örnek aday eşdizimlerine ait uzaklıkların grafiksel gösterimi

Tablo 4: *OSTAD Derlemi- Sözcük Gövdeleri için Ortalama –Varyans Yöntemi Sonuçları*

Gözlenme Sıklığı	Sözcük Gövdesi 1	Sözcük Gövdesi 2	Ortalama ($\bar{d}$)	Standart sapma ($\sqrt{s^2}$)	Gözlenme Sıklığı	Sözcük Gövdesi 1	Sözcük Gövdesi 2	Ortalama ($\bar{d}$)	Standart sapma ($\sqrt{s^2}$)
14	bilimsel	devrim	1.57	1.45	7	arı	kırlangıc	3.29	1.38
13	anne	baba	2.08	1.26	7	bakan	kurul	1.57	1.51
13	ol	ol	2.77	1.59	7	başbakan	gül	2.00	1.00
12	genel	başkan	1.00	0.00	7	başka	başka	2.57	0.98
12	naci	bey	1.00	0.00	7	el	masa	1.14	0.38
12	saime	sezgin	1.00	0.00	7	fark	et	1.00	0.00
11	beyaz	peynir	1.00	0.00	7	gözlem	deney	1.86	0.38
11	erkek	park	1.00	0.00	7	ifade	et	1.43	1.13
11	karar	ver	1.00	0.00	7	kabul	et	1.00	0.00
11	yüz	yüz	3.18	1.47	7	milli	eğitim	1.00	0.00
10	bilim	dünya	1.50	1.08	7	milyon	lira	3.00	1.00
10	kendi	kendi	1.70	1.34	7	nere	bil	1.14	0.38
10	milli	savunma	1.00	0.00	7	nesnel	gerçekliğ[3]	1.29	0.49
10	nusret	senem	1.00	0.00	7	ortak	ol	1.57	1.51
10	üvey	baba	1.50	1.08	7	rakı	sofra	1.00	0.00
9	bilim	yeni	2.78	1.30	7	sigara	sigara	2.86	0.90
9	dün	gece	1.00	0.00	7	sigara	içme	1.00	0.00
9	faiz	yüz	2.00	1.12	7	yavaş	yavaş	1.00	0.00
9	gecikme	faiz	1.78	1.20	6	a	tip	1.00	0.00
9	içeri	gir	1.33	1.00	6	alışveriş	bilim	1.17	0.41
9	taşha	kapı	1.22	0.67	6	ara	sıra	1.00	0.00
8	gül	abla	1.00	0.00	6	ara	kendi	4.00	1.67
8	kim	kim	3.50	1.51	6	ara	göz	2.67	1.37
8	masa	otur	1.00	0.00	6	arı	yuva	3.50	1.22
8	neden	ol	1.50	1.41	6	atila	sav	1.00	0.00
8	tayyip	erdoğan	1.00	0.00	6	ban[4]	ver	2.50	1.64
8	uçak	kaza	1.00	0.00	6	ban[iv]	bak	1.17	0.41
8	üst	koy	1.25	0.46	6	başbakan	yardımcı	1.00	0.00
7	adım	adım	2.71	1.70	6	başka	türlü	1.33	0.82
7	ara	ilişki	1.86	1.07	6	başka	ol	2.17	0.75

Tablo 5: *Bilkent Derlemi-Sözcük Gövdeleri için Ortalama –Varyans Yöntemi Sonuçları*

Gözlenme Sıklığı	Sözcük Gövdesi 1	Sözcük Gövdesi 2	Ortalama ($\overline{d}$)	Standart sapma ($\sqrt{s^2}$)	Gözlenme Sıklığı	Sözcük Gövdesi 1	Sözcük Gövdesi 2	Ortalama ($\overline{d}$)	Standart sapma ($\sqrt{s^2}$)
947	ol	ol	3.17	1.42	236	önem	ol	2.37	1.34
600	ol	et	3.04	1.15	234	ol	iste	2.47	1.47
519	orta	çık	1.09	0.52	229	ver	ol	3.18	1.42
483	ol	söyle	1.78	1.31	228	al	ol	3.32	1.32
468	devam	et	1.04	0.32	224	iddia	et	1.28	0.89
446	kabul	et	1.13	0.64	223	et	et	3.30	1.32
432	ol	belir	1.59	1.18	221	genel	başkan	1.44	1.14
377	türkiye	ol	3.41	1.23	216	teknik	direktör	1.04	0.38
361	yap	ol	3.11	1.37	212	iç	ol	2.61	1.48
344	ifade	et	1.20	0.78	211	ol	al	3.34	1.33
343	ol	yap	3.18	1.33	210	ara	ol	3.12	1.30
342	insan	hak	1.21	0.84	208	yap	açıkla	1.43	0.90
341	et	ol	3.04	1.45	208	ol	kendi	2.88	1.34
312	sahip	ol	1.43	1.09	208	milyar	dolar	1.36	0.86
311	dikkat	çek	1.08	0.41	208	dil	getir	1.09	0.56
303	ol	gör	1.80	1.29	199	görev	yap	1.66	1.26
294	karar	ver	1.36	0.95	198	trilyon	lira	1.25	0.86
288	neden	ol	1.86	1.31	197	ön	sür	1.08	0.52
273	konu	ol	2.42	1.41	197	tansu	çiller	1.06	0.41
267	ol	çık	2.77	1.34	196	hal	getir	1.09	0.50
265	ol	bil	1.80	1.28	195	ol	üzere	1.19	0.77
263	ol	gerek	2.33	1.50	195	kıbrıs	rum	1.29	0.90
262	ol	ver	3.40	1.22	193	dışişleri	bakan	1.09	0.42
259	milyon	dolar	1.26	0.84	190	yol	aç	1.09	0.51
248	ol	türkiye	2.55	1.40	186	başbakan	erbakan	1.57	0.73
247	genel	müdür	1.10	0.52	181	yüz	ol	3.06	1.23
244	kendi	ol	3.27	1.19	178	devlet	ol	2.84	1.34
243	ülke	ol	2.84	1.52	176	sanat	galeri	1.00	0.00
242	milyon	lira	1.57	1.08	175	resim	sergi	1.33	0.96
238	ikinci	yarı	1.02	0.23	175	karşı	çık	1.46	1.15

Ortalama-varyans yönteminde sözcük/sözcük gövde ikilileri varyans veya standart sapma değerleri azalacak şekilde listelenerek yöntemin tüm aday ikililer içinden eşdizimleri ayırt etmek konusundaki yetisi değerlendirilebilir. Ancak bu durumda çok düşük miktarda, örneğin sadece bir kez, bir arada gözlenen adayların standart sapma değeri sıfır olacaktır. Bu ikililer listenin üst sıralarında yer alırken doğru eşdizimler listenin alt sıralarında yer alacaktır. Dolayısıyla derlemdeki tüm eşdizimlerin belirlenmesinde yöntem istenilen başarım değerlerini üretemeyecektir. Bu çalışmada bir derlem veya metindeki ardışık iki sözcükten oluşan eşdizimlerin belirlenmesi amaçlanmaktadır. Bu sebeple ortalama-varyans yönteminin diğer yöntemler ile kıyaslanmasının uygun olmayacağına karar verilmiştir.

Smadja yöntemi (1993) özellikle bir anahtar sözcüğün eşdizim oluşturup oluşturmadığı ve hangi sözcük ile eşdizim oluşturduğu konularında bilgi vermektedir. Bu çalışma kapsamında amaç bir derlem veya metinde yer alan eşdizimlerin belirlenmesidir. Bu sebeple OSTAD derleminde en az 5 kere yan yana gözlenen ($f \geq 5$) gövde ikilileri güç değeri (k) azalacak şekilde listelenmiştir. Bu listede $k \geq 1$ olan ikililer Tablo'6 da verilmektedir. Tablo 6'da, k aday eşdizimin gücünü, U gözlenen sıklık değerlerinin ortalamadan ne kadar farklılaştığını, j ise aday eşdizimde sözcükler arası uzaklığı simgeler. Örneğin $j = a$ olması sözcük gövdeleri arasında $a-1$ adet sözcük yer aldığını gösterir. Bu tabloda başka sözcükler ile hiç yan yana gözlenmeyen sadece birbirleri ile gözlenen sözcük gövdelerinin güç değerleri sonsuz olduğu için liste başında yer almaktadırlar.

Çalışmada Smadja yönteminin (1993) bir veya birkaç anahtar sözcük yerine tüm derlem için uygulanarak diğer yöntemlerle kıyaslanması gereklidir. Bu durumda yöntem yüksek başarım değerleri üretirken yüksek zaman karmaşıklığı sorununu da beraberinde getirmektedir. Bu durum göz önünde bulundurularak bu yöntem diğer yöntemlerle kıyaslanmamıştır.

Tablo 6: *OSTAD Derlemi- Sözcük Gövdeleri için Smadja Yöntemi Sonuçları*

Gözlenme Sıklığı (f)	Sözcük Gövdesi 1	Sözcük Gövdesi 2	k	$U > U_0$	j	Gözlenme Sıklığı (f)	Sözcük Gövdesi 1	Sözcük Gövdesi 2	k_1	U	j
12	naci	bey	∞	23.04	1	6	devam	et	∞	5.76	1
14	bilimsel	devrim	j	21.76	1	6	dikkat	çeken	∞	5.76	1
11	beyaz	peynir	∞	19.36	1	6	dikkat	çek	∞	5.76	1
11	karar	ver	∞	19.36	1	6	günah	çıkarma	∞	5.76	1
10	nusret	senem	∞	16	1	6	orta	çıkma	∞	5.76	1
8	tayyip	erdoğan	∞	10.24	1	6	a	tip	1	5.76	1
9	içeri	gir	∞	9.76	1	5	genel-kurmay	başkan	1	4	1
9	taşha	kapı	∞	9.76	1	5	hiçbir	zaman	1	4	1
7	rakı	sofra	∞	7.84	1	5	memur	maaş	1	4	1
7	yavaş	yavaş	∞	7.84	1	5	yanlış	ortak	1	4	1
8	neden	ol	∞	7.44	1	5	gece	zaman	1	2.4	1
6	atila	sav	∞	5.76	1	5	kaza	ilgili	1	2.4	1
6	casino	venüs	∞	5.76	1	6	kadın	erkek	1	2.16	2
6	çizgi	roman	∞	5.76	1	5	uçağ	ilgili	1	1.2	3
6	çizgi	kahra-man	∞	5.76	2						

Ön değerlendirme sonucunda Smadja ve ortalama-varyans yöntemlerinin bir derlemde yer alan tüm eşdizimleri belirlemek söz konusu olduğunda maliyetlerinin yüksek olduğuna karar verilmiştir. Bu sebeple diğer yöntemler ile kıyaslanmalarının uygun olmadığı sonucuna varılmıştır.

5.2.Yöntemlerin Karşılaştırılması: Ön değerlendirme sonrasında gözlenme sıklığı, noktasal karşılıklı bilgi katsayısı, log-benzerlik, t-testi ve ki-kare testi Bilkent derleminde uygulanarak eşdizimleri belirlemekteki başarıları ölçülmüştür (Kumova Metin ve Karaoğlan 2010). Yöntemler kıyaslanırken testlerin basitleştirilmesi adına sadece ardışık iki sözcükten oluşan adayların eşdizim olup olmadığı konusundaki sonuçlar dikkate alınmıştır. Her yöntemin Bilkent derlemi içinde yer alan ikilileri eşdizim olma ihtimallerini değerlendirerek sıralamaları sağlanmıştır. Daha sonra her yöntemin ilk 200 adayı seçilerek temel veri kümeleri oluşturulmuştur[5]. Sonuçta sözcük gövdelerinden oluşan 661 aday ikili, gövdelenmemiş sözcüklerden oluşan 506 aday ikili ile iki temel veri kümesi oluşmuştur.

Eşdizim belirleme yöntemlerinin değerlendirilmesi ve kıyaslanmasında anma (*recall*) ve duyarlılık (*precision*) ölçütlerinin bir arada değerlendirildiği F-ölçütü kullanılmıştır. Burada anma değeri, r, bir yöntemin temel veri kümesi içinde doğru eşdizim olarak aday gösterdiği sözcük ikililerinin temel veri kümesi içinde yer alan tüm doğru eşdizimlere oranı olarak tanımlanabilir. Duyarlılık değeri, p, ise bir yöntemin eşdizim olarak aday gösterdiği ikililer içindeki doğru eşdizimlerin aday ikililere oranı olarak ifade edilir. Bu iki ölçeğin harmonik ortalaması olan F-ölçütü (F-değeri) şu şekilde hesaplanır:

$$F = 2 \cdot \frac{p \cdot r}{p + r} \qquad (11)$$

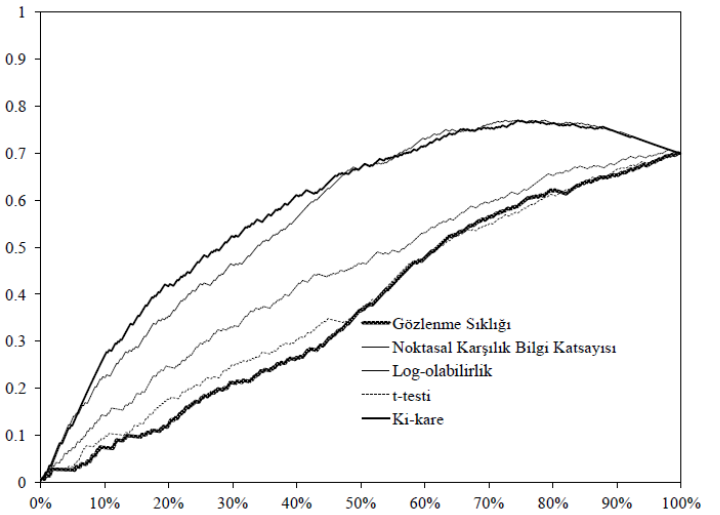
Çalışmamızda F-değerinin sunulmasında Evert ve Krenn çalışmasında (2001) izlenen yöntem tercih edilmiştir. Bu yaklaşımda, her bir yöntemin temel veri kümesinde yer alan ikilileri eşdizimlilik ihtimali azalacak şekilde sıralayarak bir liste oluşturması sağlanır. Bu listenin üst sıralarında yöntemin güçlü eşdizim adayları, alt sıralarında ise eşdizimlilik ihtimali zayıf adayları bulunur. Duyarlılık ve anma değerleri ilk N aday göz önünde bulundurularak hesaplanır. N değeri 1'den başlayarak liste uzunluğu olan n değerine ulaşıncaya kadar birer birer arttırılarak n adet duyarlılık ve anma değeri elde edilir. Bu değerler denklem 11'de yerine konarak her bir adımda F-değeri hesaplanır. Bu yaklaşımda, bir yöntemin başarımına tek bir N değerinde bakmak yerine (örneğin $N = 1$ veya $N = n$ gibi) ürettiği listedeki başarımın ne şekilde değiştiğine n adet değerden oluşan F-eğrisi ile bakılır. Tüm F-eğrileri belirli bir taban F-değeri ile sonlanır. Bu taban değer, duyarlılık taban değerine bağlıdır. Herhangi bir yöntem temel veri kümesinde yer alan tüm ikilileri sıraladığında ($T(w_i w_j)$ iken) tüm doğru eşdizimleri de belirli sıralara atamış olmak zorundadır, yani anma değeri $r = 1$ olmuştur. Bu durumda duyarlılık değeri ise tüm kümenin doğru eşdizim oranına ulaşmıştır. Bu sebeple, örneğin %53.5 oranında doğru eşdizim içeren bir veri kümesinde taban F-değeri

$2 \cdot \frac{0.535}{0.535 + 1} = 0.697$ olacaktır.

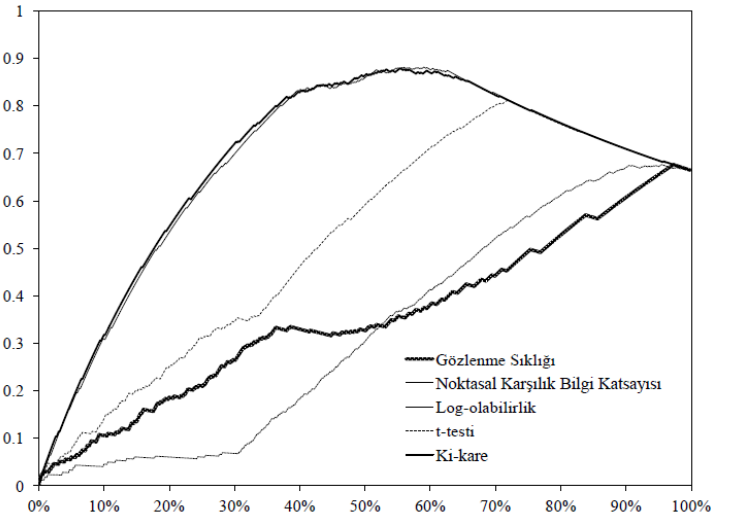
Temel veri kümelerinden elde edilen F-değer grafikleri Şekil 2 ve 3'de verilmiştir. Bu şekillerde yatay eksen N / n ($N = 1 \ldots n$) oranının yüzde olarak

ifadesi, dikey eksen ise ilgili andaki F-değeridir. F eğrilerinde başarı göstergesi temel veri kümesinde bulunan doğru eşdizimlerin ilk sıralarda yakalanması sonucunda ilgili eğrilerin grafiğin sol üst köşesine yaklaşmasıdır. Aynı temel veri kümesi sözkonusu olduğu için tüm eğriler grafiğin sağ kısmında birbirine yaklaşmış ve taban F-değerinde sonlanmışlardır. Sözcük ikilileri içeren temel veri kümesi için taban değer=0.697 (Şekil 2), gövde ikilileri içeren temel veri kümesi için taban değer=0.665'dir (Şekil 3).

Şekil 2 ve 3'de verilen F-değer grafiklerinde 3 temel özellik dikkat çekmektedir. İlk özellik grafiklerin genelinde χ^2 (ki-kare) ve noktasal karşılıklı bilgi katsayısı yöntemlerinin diğer yöntemlere oranla daha yüksek F-değerlerine sahip olmasıdır. Log-olabilirlik, t-testi ve gözlenme sıklığı yöntemleri ise eğrilerin büyük bir kısmında taban F-değerinin altında kalmışlardır. Gözlenme sıklığı yöntemi sözcüklerin gövdelenmediği durumda diğer dillerdeki çalışmalarda da olduğu üzere tüm yöntemlere oranla daha başarısız olmuştur. İkinci özellik, gövde ikililerinden oluşan temel veri kümesinde yöntemlerin diğer veri kümesine oranla daha yüksek başarı göstermeleridir. Bu özellik, istatistiksel eşdizim belirlemede başarımın gövdeleme ile arttığına dair bir gösterge olarak kabul edilebilir. F-değer grafikleri incelendiğinde ortaya çıkan üçüncü özellik, gövde ikililerinden oluşan temel veri kümesinin başarılı ve başarısız olarak gruplayabileceğimiz yöntemleri birbirinden ayırmada daha net sonuçlar ürettiğidir. Bu veri kümesinde en başarılı yöntemler (χ^2 ve noktasal karşılıklı bilgi katsayısı) başarım açısından birbirlerine çok yaklaşırken diğer yöntemler başarım açısından arayı oldukça açmaktadırlar.



Şekil 2. Bilkent derlemi- Sözcük ikilileri içeren temel veri kümesinde F-değer grafiği.

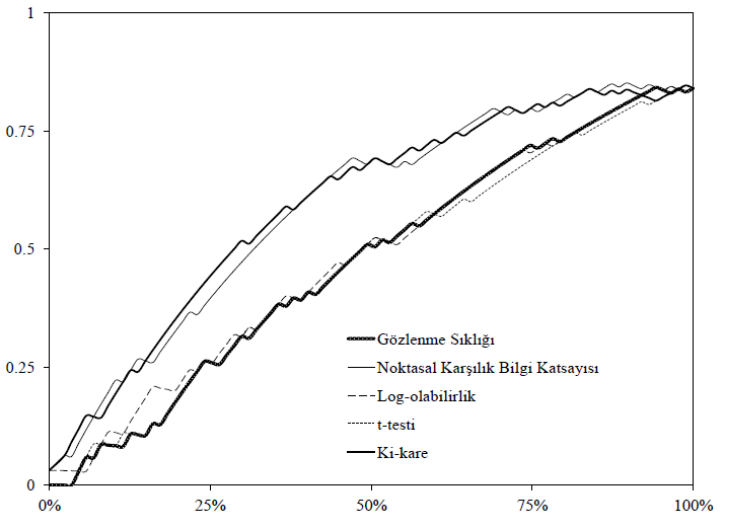


Şekil 3. Bilkent derlemi-Gövde ikilileri içeren temel veri kümesinde F-değer grafiği.

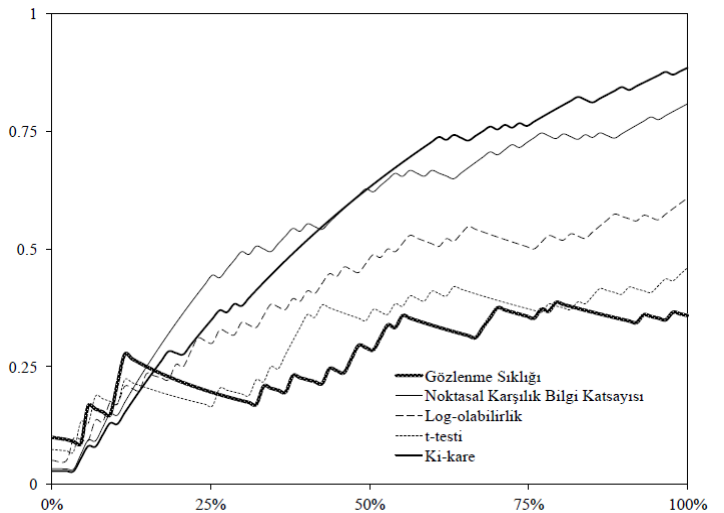
5.3.Eşdizim Eğilimi Yönteminin Uygulanması ve Değerlendirilmesi: Eşdizim eğilimi yönteminin sınanması için Bilkent derlemi kullanılmıştır. Gövdelenmemiş sözcükler üzerinde gözlenme sıklığı, noktasal karşılıklı bilgi katsayısı, log-benzerlik, t-testi ve ki-kare testleri uygulanarak temel veri kümesi elde edilmiştir. Bu veri kümesi üzerinde tüm küme için en yüksek duyarlılık değerini (0.724) üreten $T_0 = (0.2\ )^{-1}$ eşdizim eğilimi yöntemi için eşik değer olarak kabul edilmiştir. Yöntemin uygulanması sonucunda 506 aday ikilinin yer aldığı temel veri kümesi 87 ikili içeren bir liste haline gelmiştir.

Bu aday listesi her bir yönteme göre sıralanmış ve F-değer grafikleri hazırlanmıştır. Şekil 4 eşdizim eğilimi yöntemi sonrasında elde edilen F-değer eğrilerini içermektedir. Eşdizim eğilimi yönteminin sonuçlarının aynı sayıda aday içeren liste ile kıyaslanması amacıyla bu yöntem uygulanmadan önce yöntemlerin ürettiği ilk 87 adaya ait F-değer grafiği de Şekil 5'de verilmiştir. Şekil 4 ve 5'de yatay eksen tüm veri kümesinin ne oranda tamamlandığını ($N = 1 \ldots n$ iken N / n oranının yüzde olarak ifadesi) gösterir, dikey eksen ise ilgili andaki F-değerini verir.

Şekil 4 ve 5'de yer alan F-değer grafikleri kıyaslandığında eşdizim eğilimi yönteminin uygulanması sonucunda yöntemlerde farklı oranlarda iyileşme görülmektedir. Özellikle t-testi, log-olabilirlik ve gözlenme sıklığı yöntemleri diğer yöntemlere yaklaşacak oranda iyileşme göstermektedir. Bu sayede eşdizimlerin belirlenmesinde en basit yöntem olarak kabul edilen gözlenme sıklığı yöntemi diğer yöntemler ile yarışabilir hale gelmişttir. Ayrıca F-değer eğrileri incelendiğinde tüm yöntemlerin F-değerlerinin (eğrilerinin) eşdizim eğilimi yöntemi ile birbirine yaklaştığı görülmektedir. Bu sonuçlar eşdizim eğilimi yönteminin Türkçe metinler üzerinde uygulanabilir olduğunu dolayısıyla yöntemin temelinde yatan anlam bütünlüğü fikrinin geçerliliğini göstermektedir.



Şekil 4. Eşdizim eğilimi yönteminin uygulanması sonucu elde edilen F-değer grafiği.



Şekil 5. Eşdizim eğilimi yönteminin uygulanması öncesinde yöntemlere ait ilk 87 adaydan elde edilen F-değer eğrileri.

Eşdizimlilik ölçümünde istatistiksel yöntemlerin etkinliğinin sergilendiği bu çalışmadan yola çıkarak ileride bu yöntemlerin iyileştirilmesi ve Türkçe'deki sözcükler arası bağların kuvvetini ölçmeye dayalı çalışmalar yürütülmesi planlanmaktadır. Ayrıca çalışmamızda eşdizimlilik olarak nitelendirilen sözcük birlikteliklerini yaratan zorunlu morfolojik yapıların belirlenmesi bir diğer araştırma konusu olarak planlanmıştır.

Açıklamalar

1 Türkiye Türkçesinde eşdizimliliğin kavramsal tartışması için Özkan (2007), Özkan (2010), Mersinli ve Demirhan (2012) kaynaklarından faydalanılabilir.

2 Hipotez 1 ve 2 için verilen eşitliklerin ilk terimleri ($b(c_{12}, c_1, p)$ ve $b(c_{12}, c_1, p_1)$) w_2 sözcüğünün derlemde c_1 kere rastlanan $\sqrt{s^2}$ sözcüğü ile bir arada bulunduğu c_{12} adet ikili olmasına dair toplam olasılığı simgeler. İkinci terimlerinde ($b(c_2 - c_{12}, N - c_1, p)$ ve $b(c_2 - c_{12}, N - c_1, p_2)$) ise w_1 sözcüğünün bulunmadığı N-c_1 adet ikilide w_2 sözcüğünün c_2-c_{12} kere bulunmasına dair olasılık hesabı yer alır. Hipotezlerde bu iki durumun bir arada gerçekleşmesi gerektiği için iki duruma ait bileşik olasılık hesaplanmaktadır.

3 Bu sözcük, kullanılan derlemde "gerçekliği" sözcüğünün gövdesi olarak verilmiştir. Derlemde yer alan bu tip gövdeleme hataları sözcük sıklık değerlerini etkilememek adına çalışmamızda düzeltilmemiştir.

4 Bu sözcük, kullanılan derlemde "bana" sözcüğünün gövdesi olarak verilmiştir. Derlemde yer alan bu tip gövdeleme hataları sözcük sıklık değerlerini etkilememek adına çalışmamızda düzeltilmemiştir.

5 Log-olabilirlik, ki-kare (serbestlik derecesi=1) ve t-testi (serbestlik derecesi > 1000) için ilk 200 aday seçilirken $\alpha = 0.005$ kabul edilmiştir.

Kaynaklar

Aksan, Yeşim (2011). "Derlem temelli sözcük anlambilimi çalışmalarının Türkçenin eğitimine katkısı" *Theoretical and Applied Researches in Turkish Language Teaching (L. Uzun & Ü. Bozkurt), 345-358*. Essen: Die Blaue Eule.

Atalay, N. Bedin vd. (2003). "The Annotation Process in the Turkish

Treebank*" Proceedings of the EACL Workshop on Linguistically Interpreted Corpora–LINC*. Budapest, Hungary.

Baker, Paul vd. (2006). *A Glossary of Corpus Linguistics*. Edinburg University Press.

Bisht, R. Kishore vd. (2006). "An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations*". Journal of Quantitative Linguistics*(13): 161-175.

Church, K. Ward ve Hanks, Patrick (1990). "Word Association Norms, Mutual Information, and Lexicography*" Computational Linguistics*(16): 22-29.

Dinçer, Taner (2004). *Türkçe için istatistiksel bir bilgi geri-getirim sistemi, Doktora Tezi.* U.B.E., Ege Üniversitesi.

Dunning, Ted (1993). *"Accurate* methods for the statistics of surprise and coincidence*". Computational Linguistics* (19): 61–74.

Evert, Steven ve Krenn, Brigette (2001*).* "Methods for the qualitative evaluation of lexical association measures*" Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Toulouse, Fransa.

Firth, John Rupert (1957). "Modes of Meaning*". Papers in Linguistics 1934-51*. Oxford University Press.

Hartmann, Reinhard Rudolf Karl ve James, Gregory (1998). *Dictionary of Lexicography*. London: Routledge.

Hindle, Donald (1990). "Noun Classification from Predicate-Argument Structures*" Annual Meeting of the Association for Computational Linguistics (ACL 1990.*, Pittsburgh, Pennsylvania, ABD.

Hoey, Michael (1991) *Patterns of Lexis in Text*. Oxford University Press.

Justeson, John S. ve Katz, Slava M. (1995). "Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns", *Computational Linguistics* (21).

Kita, Kenji vd. (1994). "A comparative study of automatic extraction of

collocations from corpora: Mutual information vs. cost criteria", *Journal of Natural Language Processing*(1): 21-33.

Kumova Metin, Senem ve Karaoğlan Bahar (2010). "Collocation Extraction in Turkish Texts Using Statistical Methods" *7th International Conference on Natural Language Processing (LNCS-ISI) IceTAL 2010*. Reykjavik, Iceland.

____ (2011). "Measuring Collocation Tendency of Words" *Journal of Quantitative Linguistics* (18):174-187.

Malmkjaer, Kirsten (2001). *Linguistics Encyclopedia*. Florence. KY. USA: Routledge.

Manning, Chris D. ve Schütze, Hinrich (1999) *Foundations of Statistical Natural Language Processing*. The MIT Press.

Mersinli, Ümit ve Demirhan, Umut (2012). "Çok Sözcüklü Kullanımlar ve İlköğretim Türkçe Ders Kitapları" *Türkçe Öğretiminde Güncel Çalışmalar (Aksan, M. ve Aksan, Y.)*:113-122. Mersin: Mersin Üniversitesi

Oflazer, Kemal vd. (2004). "Integrating Morphology with Multi-word Expression Processing in Turkish", *2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004).* Barcelona, İspanya.

Özkan, Bülent (2007). *Türkiye Türkçesinde Belirteçlerin Fiillerle Birliktelik Kullanımları ve Eşdizimliliği*. Doktora Tezi, Çukurova Universitesi, Adana.

____ (2010). "Türkçenin Öğretiminde Sıfatların Eşdizim Sözlüğü: Yöntem ve Uygulama" *e-International Journal of Educational Research* (1: 51-65).

____ (2012). "Türkiye Türkçesinin Eşdizim Sözlüğü" *IV. Uluslararası Dünya Dili Türkçe Sempozyumu:*93-102. Muğla/Türkiye.

Pearce, Darren (2002). "A comparative evaluation of collocation extraction techniques" *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Spain.

Sarıkaş, Ferah (2006). "Problems in Translating Collocations", *Elektronik Sosyal Bilimler Dergisi* (5):33-40.

Shimohata Sayori vd. (1997). "Retrieving collocations by co-occurrences and word order constraints", *The Eighth Conference on European Chapter of the Association for Computational Linguistics*. Madrid, İspanya.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Smadja, Frank (1993). "Retrieving Collocations from Text: Xtract" *Computational Linguistics*(19): 143 – 177.

Sterkenburg, Piet Van (2003). A Practical Guide to Lexicography. Amsterdan/Philedelphia: John Benjamins Publising Company.

Taşıgüzel, Selver (1988). "İlköğretim Türkçe Ders Kitaplarında Öğretici Nitelikli Metinlerdeki Eşdizimsel Örüntülerin Görünümü", *Dil Dergisi*. Ankara Üniversitesi Türkçe ve Yabancı Dil Araştırma ve Uygulama Merkezi.

Tür, Gökhan vd. (2003). "A Statistical Information Extraction System for Turkish" *Natural Language Engineering*(9):181-210.

Van Buren, Paul (1967). "Preliminary Aspects of Mechanisation in Lexis" *Cahiers de Lexicology, 89-112, 12 71-84.*

Identifying Collocations in Turkish Using Statistical Methods

Senem Kumova Metin*

Bahar Karaoğlan**

Abstract

Collocation is the combination of words in which words appear together more often than by chance in order to create a block of meaning. Since the extraction of collocations provides many benefits in automatic processing, translation of Turkish texts and in learning Turkish, it is an important issue in Turkish natural language processing. In this study several statistical techniques, including occurrence frequency, pointwise mutual information and hypothesis tests, are applied on Turkey Turkish corpus to automatically identify collocations. We have utilized both stemmed and surface forms of words in order to explore the effect of stemming in collocation extraction. The techniques are evaluated using the F-measure. The chi-square hypothesis test and pointwise mutual information methods have produced better results compared to other methods. In addition, we have observed that when words are stemmed, methods which may be considered as successful in collocation extraction may be more clearly discriminated.

Keywords

Collocation, Turkey Turkish, natural language processing, corpus

* Assist. Prof.Dr., İzmir University of Economics, Faculty of Engineering and Computer Science, Department of Software Engineering – İzmir/Turkey
senem.kumova@ieu.edu.tr

** Prof.Dr., Ege University, International Computer Institute – İzmir/Turkey
bahar.karaoglan@ege.edu.tr

Определение словосочетание в турецком языке с использованием статистических методов

Сенем Кумова Метин*

Бахар Караоглан**

Аннотация

Словосочетание-это сочетание слов, в которой вместе эти слова встречаются гораздо чаще, чем в случайном порядке по отдельности, чтобы создать определенное значение таким образом, выявление таких коллоквизмов дает массу преимуществ в автоматической обработке данных в переводах турецкого текста и в изучении турецкого языка. Это важный аспект в естественном изменение турецкого языка. В этом исследовании применяются некоторые статические техники, частота повтора, тесты взаимно важной информации и гипотетические тесты. В турецком языке, чтобы определить эффект основы в выборе коллоквизмов.

Эти техники оцениваются по F- шкале. Гипотический тест «Квадрат Чи» и метод взаимо важной информации обеспечили лучшие результаты по сравнению с другими методами. Более того, мы пришли к выводу, что в словосочетаниях, где слова «насажены», эти методы, которые могут считаться успешными с коллоквизмами, будут не столь полезны.

Ключевые слова

словосочетание, турецкий язык, естественное изменение языка

* и.о.доц.док., Университет Измир Экономика, факультет Инженерия и комьпютерная технология, кафедра Програмного обеспечение–Измир /Турция
senem.kumova@ieu.edu.tr