

EGE ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

(DOKTORA TEZİ)

**TÜRKÇEDE HESAPLAMALI
METİN ANALİZİ**

Senem KUMOVA METİN

Tez Danışmanı : Prof. Dr. Bahar KARAOĞLAN

Uluslararası Bilgisayar Anabilim Dalı

Bilim Dalı Kodu : 619.02.04

Sunuş Tarihi : 17.02.2011

**Bornova-İZMİR
2011**

Senem KUMOVA METİN tarafından **doktora tezi** olarak sunulan “**Türkçede Hesaplamalı Metin Analizi**” başlıklı bu çalışma E.Ü. Lisansüstü Eğitim ve Öğretim Yönetmeliği ile E.Ü. Fen Bilimleri Enstitüsü Eğitim ve Öğretim Yönergesi’nin ilgili hükümleri uyarınca tarafımızdan değerlendirilerek savunmaya değer bulunmuş ve 17.02.2011 tarihinde yapılan tez savunma sınavında aday oybirliği/oyçokluğu ile başarılı bulunmuştur.

Jüri Üyeleri:

İmza

Jüri Başkanı	: Prof. Dr. Bahar Karaoğlu
Raportör Üye	: Yrd. Doç. Dr. Cengiz Güngör
Üye	: Yrd. Doç. Dr. B. Taner Dinçer
Üye	: Prof. Dr. Aydın Öztürk
Üye	: Yrd. Doç. Dr. Muhammed Cinsdikici

ÖZET**TÜRKÇEDE HESAPLAMALI
METİN ANALİZİ**

KUMOVA METİN, Senem

Doktora Tezi, Uluslararası Bilgisayar Enstitüsü.

Tez Danışmanı: Prof. Dr. Bahar Karaoğlu

Şubat 2011, 81 sayfa

Metinler bilginin dil vasıtasıyla yazılı biçimde aktarımını ve saklanmasını sağlayan araçlardır. Bu tezde metnin taşıdığı anlamın, aktarmak istediği enformasyonun tespiti çalışmalarını destekleyen iki temel analiz çalışması sunulmaktadır: işlev kelimelerin ve eşdizimlerin tespiti.

İşlev kelimeler metnin anlam bütünlüğüne katkısı olmayan ya da çok az olan kelimelerdir. Bu kelimeler sıklıkla gözlenmeleri sebebiyle metinde azımsanamayacak büyüklükte bir yer işgal ederler. Tez kapsamında işlev kelimelerin tespiti için dile ait farklı özelliklerin bir arada değerlendirilebileceği istatistiksel bir yöntem olan diskriminant analizinin kullanılması önerilmiştir. Önerilen yöntem Türkçe metinler üzerinde sınanmış ve sıklıkla kullanılan gözlenme sıklığı ve tf-idf gibi mevcut yöntemlerden daha başarılı sonuçlar ürettiği gösterilmiştir.

Eşdizimler anlam bütünlüklerini tamamlamak veya güçlendirmek amacıyla birarada bulunan kelimelerin oluşturduğu kelime birlikleridir. Eşdizim oluşturan kelimelerin birbirlerini anımsattığı temelinden yola çıkarak tez kapsamında eşdizim eğilimi yöntemi geliştirilmiştir. Öncelikle literatürde yer alan bazı yöntemler Türkçe derlemeler üzerinde uygulanmış ve gövdelemenin etkisi araştırılmıştır. Daha sonra önerilen yöntem mevcut yöntemlerin ürettiği temel veri kümesine tatbik edilerek, bu yöntemlerin duyarlılık ve anma değerlerindeki değişim incelenmiştir. Eşdizim eğilimi yönteminin tüm yöntemleri iyileştirdiği gözlenmiştir.

Anahtar Kelimeler: Metin analizi, işlev-içerik kelime, eşdizim.

ABSTRACT**COMPUATIONAL TEXT ANALYSIS
IN TURKISH**

KUMOVA METİN, Senem

PhD in International Computer Institute.

Supervisor: Prof. Dr. Bahar Karaoğlu

February 2011, 81 pages

Texts are language tools to transfer and store information in written form. In this thesis, two analysis methods: function word detection and collocation extraction which support the studies in extraction of meaning and measuring the amount of information are presented.

Function words, playing grammatical roles, appear in high frequencies and rarely contribute to the meaning of the text. In this thesis, a statistical method, discriminant analysis, which merges several linguistic features, is proposed to discriminate function words. The proposed method is utilized on Turkish texts. The results show that the method brings a significant improvement in distinction compared to well known frequency of occurrence and tf-idf methods.

Collocations are the groups of words which collocate to complete or enforce their meaning integrity. In this thesis, collocative tendency method which states that any word in a collocation must suggest or at least imply the following words composing the collocation is presented. Firstly, current collocation extraction methods are utilized on a Turkish corpora and the effect of stemming is investigated. Following, the proposed method is tested on a base data set extracted by some statistical techniques and it is evaluated by precision and recall measures. It is found that collocation tendency method gives a remarkable improvement in current techniques of collocation extraction.

Keywords: Text analysis, function-content word, collocation.

TEŞEKKÜR

Öncelikle bu tez konusu üzerinde bana çalışma imkânı sunan ve çalışma sırasında bilimsel katkıları ile bana yardımcı olan, eğitimim süresince yardımlarını esirgemeyen, tez danışmanım ve değerli hocam Prof. Dr. Bahar KARAOĞLAN'a en içten teşekkür ve saygılarımı sunarım.

Tez izleme toplantılarında jüri olarak görev alan Prof. Dr. Aydın ÖZTÜRK, Yrd. Doç. Dr. Bekir Taner DİNÇER'e, Yrd. Doç. Dr. Cengiz GÜNGÖR ve Yrd. Doç. Dr. Muhammed CİNSDİKİCİ'ye değerli önerileri ve çalışmaya olan katkıları için teşekkürlerimi sunarım

Doktora süresince bana desteklerini hiçbir zaman esirgemeyen başta Prof. Dr. Cemali DİNÇER olmak üzere İzmir Ekonomi Üniversitesi Mühendislik ve Bilgisayar Bilimleri Fakültesinde görev yapan değerli hocalarım ve arkadaşlarıma teşekkürlerimi sunarım.

Bu uzun çalışma süresi boyunca her zaman yanımda olan ve benden desteklerini esirgemeyen babam M. Tacettin Kumova, annem Müjgan Kumova, ablam F. Ayça Kumova, eşim Burak Metin ve biricik oğlum Selim Metin'e teşekkürü bir borç bilirim.

İÇİNDEKİLER

Sayfa

ÖZET	v
ABSTRACT	vii
TEŞEKKÜR	ix
İÇİNDEKİLER.....	xi
ŞEKİLLER DİZİNİ	xv
ÇİZELGELER DİZİNİ.....	xvii
1. GİRİŞ	1
2. METİN ANALİZİ	5
2.1. Tanımlar.....	5
2.2. İlk Çalışmalar	9
3. İŞLEV KELİMELERİN TESPİTİ	17
3.1. İşlev Kelime Tespiti Problemi	17
3.2. Literatür	19
3.3. Diskriminant Analizinin Temelleri.....	22
3.4. İşlev Kelime Tespitinde Diskriminant Analizinin Uygulanması	26
3.4.1. İşlev kelime tespitinde kullanılacak özellikler	26
3.4.2. Geçerlilik testi.....	29

İÇİNDEKİLER(devam)

Sayfa

3.4.3. Talim ve Sınama	30
3.4.4. Değerlendirme.....	32
3.5. Sonuçlar	33
4. EŞDİZİMLERİN TESPİTİ.....	37
4.1. Eşdizimlerin Tespiti Problemi	37
4.2. Eşdizim Özellikleri	39
4.3. Eşdizim Tespit Yöntemleri	41
4.3.1. Gözlenme sıklığı	41
4.3.2. Karşılıklı enformasyon.....	42
4.3.3. Hipotez testleri	43
4.3.4. Ortalama –varyans yöntemi	46
4.3.5. Smadja yöntemi.....	47
4.3.6. Shimohata - Sugio - Nagata çalışması	49
4.4. Türkçe’de Eşdizim Çalışmaları.....	52
4.5. Eşdizim Eğilimi Yöntemi	53
4.6. Deney Derlemleri.....	55

İÇİNDEKİLER(devam)

Sayfa

4.7. Deneysel Hazırlık	56
4.7.1. Temel veri kümesi	56
4.7.2. Değerlendirme yöntemi	57
4.8. Sonuçlar	57
4.8.1. Ön değerlendirme	58
4.8.2. Yöntemlerin kıyaslanması	62
4.8.3. Eşdizim eğilimi yönteminin uygulanması ve değerlendirilmesi	65
5. SONUÇ	69
KAYNAKLAR DİZİNİ.....	73
ÖZGEÇMİŞ.....	81
EKLER.....	
Ek 1 İşlev kelimelerin tespiti için kullanılan baz işlev kelime listesi	
Ek 2 Diskriminant analizinin deney belgelerinde ürettiği ilk 50 aday işlev kelime..	
Ek 3 Kelime sıklık yönteminin deney belgelerinde ürettiği ilk 50 aday işlev kelime	
Ek 4 tf-idf yönteminin deney belgelerinde ürettiği ilk 50 aday işlev kelime.....	

ŞEKİLLER DİZİNİ

<u>Şekil</u>	<u>Sayfa</u>
2.1 OSTAD derleminin gövde formu için gözlenme sıklığı – sıra numarası grafiği	11
2.2 OSTAD derleminin yüzeysel formu için gözlenme sıklığı – sıra numarası grafiği	11
2.3 Kelime gözlenme sıklık-sıra numarası eğrilerinde önemli kelimelerin ifade gücünün şematik gösterimi	14
3.1 Diskriminant analizi sonucunda elde edilen grup dağılımlarının ve ortalama değerlerinin hata bölgeleri ile beraber şematik gösterimi.	24
3.2 Diskriminant analizi, gözlenme sıklığı ve tf-idf yöntemleri için deney belgelerinden elde edilen duyarlılık grafikleri.	36
4.1. OSTAD derlemi örnek aday eşdizimlerine ait uzaklıkların grafiksel gösterimi.....	59
4.2. Bilkent derlemi için duyarlılık grafiği (yüzeysel form).....	63
4.3. Bilkent derlemi için duyarlılık grafiği (gövde formu).	63
4.4. Bilkent derlemi için anma grafiği (yüzeysel form).	64
4.5. Bilkent derlemi için anma grafiği (gövde formu).....	64
4.6 Eşdizim eğilimi yönteminin uygulanması sonucu elde edilen duyarlılık grafiği	65
4.7 Eşdizim eğilimi yöntemi öncesinde diğer yöntemlerin ürettiği ilk (en iyi) 87 aday ikili için duyarlılık grafiği	66
4.8. Eşdizim eğilimi yönteminin uygulanması sonucu elde edilen anma grafiği	67

ÇİZELGELER DİZİNİ

<u>Çizelge</u>	<u>Sayfa</u>
3.1 OSTAD derleminde sıklıkla yer alan kelime ikilileri.....	29
3.2 OSTAD derleminin genel özellikleri.....	31
3.3 İşlev-içerik kelime ayrımı yapılan örnek Türkçe metinlerin genel özellikleri.....	32
3.4 K- katlı çapraz geçerlilik testinden kullanılan gruplara ait istatistikler.	33
3.5 Talim derleminden elde edilen diskriminant analizi sonuçları.....	34
3.6 Sınama derleminde diskriminant analizi sonucunda elde edilen aday işlev kelime miktarları.....	34
3.7 Sınama belgelerinde diskriminant analizi sonucunda elde edilen duyarlılık değerleri	35
4.1 “beyaz” ve “saray” kelimeleri için gözlenen sıklık değerlerini içeren 2x2’lik tablo	44
4.2 Smadja yöntemi, örnek ikililer	48
4.3 OSTAD ve Bilkent derlemlerinin genel özellikleri	56
4.4 OSTAD derleminden seçilen örnek eşdizim adayları için birarada görünme sıklık ve uzaklık değerleri	59
4.5 OSTAD derlemi ortalama –varyans yöntemi	60
4.6 Bilkent derlemi ortalama –varyans yöntemi.....	61
4.7 OSTAD derlemi Smadja yöntemi sonuçları	62

1. GİRİŞ

Metin analizi, yazılı metinlerde aktarılmak istenen enformasyonun ve metin içinde taşınan anlamın ortaya konmasını amaçlamaktadır. Metin içinde yer alan en küçük anlamlı birim sözcüklerdir/kelimelerdir. Sözcükler bir takım ilişki ağları ile birbirlerine bağlanarak diğer anlamlı birimleri ve sonuçta metinleri oluşturarak enformasyonun akışını sağlarlar. Bu tez anlam taşıyan en küçük birim olan sözcükler vasıtasıyla metin içindeki anlamın tespitini destekleyen iki temel çalışmayı sunmaktadır: işlev kelimelerin ve iki kelimeden oluşan eşdizimlerin belirlenmesi.

Sözcükler en genel tanımıyla anlamı veya görevi bulunan ve tek başına kullanılabilen ses veya sesler topluluğudur. Ancak diller içinde yer alan bazı sözcükler ya hiç anlam taşımaz ya da taşıdığı anlam miktarının metne katkısı çok azdır. Bu sebeple metin içinde yer alan kelimeler taşıdıkları anlam miktarına göre iki grupta değerlendirilebilir: işlev kelimeler, içerik kelimeler. İşlev kelimeler, anlam taşımayan, cümle yapısının gramer kurallarına göre oluşmasını ve akıcılığı sağlayan sözcüklerdir. İçerik kelimeler ise metne anlam katkısı bulunan, metnin taşıdığı enformasyonu arttıran kelimelerdir. İşlev kelimelerin tespiti aynı zamanda içerik kelimelerin tespiti çalışmalarını ve sonuç olarak bilgi geri getirim sistemlerinde kullanılan indeks kelimelerin belirlenmesi çalışmalarını desteklemesi sebebiyle önemlidir. Tez kapsamında işlev kelimelerin tespiti için çeşitli bağımsız değişkenler vasıtasıyla bağımlı değişkenin (ilgili sözcük) farklı gruplara (işlev veya içerik kelime grubu) atanmasına olanak tanıyan diskriminant analizi yöntemi kullanılmıştır. Yöntem literatürde sıklıkla kullanılan yöntemlere (kelime sıklık yöntemi ve tf-idf) oranla kelime gruplarını ayırt etmede daha başarılı sonuçlar vermiştir.

Eşdizim, iki ya da daha çok kelimeden oluşan, kelimelerin anlamlarını tam olarak yitirmedeği/değiştirmedeği ancak bir arada kullanılmaları sonucu yeni bir anlam kazandıkları kelime gruplarıdır. Bu kelime gruplarının temel özelliği diğer kelime kombinasyonları ile kıyaslandığında daha sık gözlenmeleridir. Herhangi bir eşdizimde yer alan bir kelime eşdizim içinde yer alan diğer kelimelerden bağımsız olarak değerlendirilirse anlamsal yanlışlıklar yapılması muhtemeldir.

Örneğin “Bu adam tam bir kara cahil.” cümlesinde yer alan “kara cahil” bir eşdizimdir. Eşdizim içinde yer alan kelimeler teker teker değerlendirildiğinde siyah/kara renkli bir cahil adam anlamı ortaya çıkar oysaki bu iki kelime bir araya gelerek “çok cahil” anlamını oluşturmaktadır. Tez içerisinde ilk olarak eşdizimlerin tespitine yönelik olarak literatürde yer alan yöntemler Türkçe metinler üzerinde uygulanarak sonuçlar değerlendirilmiştir. Ardından eşdizimlerin tespiti için geliştirilen “eşdizim eğilimi” yöntemi uygulanmıştır. Eşdizim eğilimi yöntemi bazı kelimelerin anlam bütünlüklerini tamamlamak amacıyla başka kelimelere ihtiyaç duyduğu ve bu sebeple eşdizim oluşturma eğilimleri olduğu fikrine dayanmaktadır. Yapılan sınamalar sonucunda yöntemin mevcut yöntemlerin başarımlarını arttırdığı gözlenmiştir.

Tez kapsamında geliştirilen ve farklı bir amaçla kullanılması önerilen mevcut yöntemler Türkçe metinler üzerinde test edilmiştir. Ancak yöntemlerin tümü aslen dil bağımsızdır yani farklı diller için de aynı yöntemlerin uygulanabilirliği söz konusudur.

Tez içerisinde gerçekleştirilen çalışmalar doğal dil işleme ana başlığı altında makine çevirisi, otomatik özetleme, doğal dil anlama, kelime anlam belirsizliklerinin giderilmesi vb. gibi pek çok farklı alana ek olarak bilgi geri getirim, indeksleme gibi anlamsal enformasyon ölçümü gerektiren yapay zeka çalışmalarına da hizmet etmektedir.

Tezin ana bilgi gövdesine katkıları iki kısımda incelenebilir. İşlev kelime tespiti çalışması sonucu ortaya çıkan katkılar şu şekilde listelenebilir:

- Literatürde, yazılı metinlerde işlev kelimelerin tespiti için gözlenme sıklığı temeline dayanan çalışmalar dışında kelime ve dil özelliklerini de kullanarak otomatik listelemeyi sağlayan bir yöntem bilginiz dâhilinde yoktur. Önerilen diskriminant analizi yöntemi işlev kelimelerin tespiti için farklı metin özelliklerini bir arada değerlendirmesi bakımından bir ilktir.
- İşlev kelime tespiti için önerilen yöntem sadece Türkçe için değil farklı dillerde de uygulanabilir niteliktedir.

- İşlev kelimelerin otomatik tespitinde hâlihazırda sıklıkla kullanılan gözlenme sıklığı ve tf-idf yöntemleri işlev kelime adaylarını sıralı bir liste halinde sunmaktadırlar. Bu sebeple belirli bir eşik değeri belirlenmesi sorunu ortaya çıkmaktadır. Araştırmacılar tarafından belirlenen bu eşik değeri sınıflama başarımını doğrudan etkilemektedir. Diskriminant analizinde ise derleme özgü işlev kelime listesi eşik değeri problemi olmadan ortaya çıkmaktadır.

Metinlerdeki eşdizimlerin belirlenmesi çalışmasının katkıları ise şu şekilde listelenebilir:

- Literatürde yer alan çeşitli eşdizim tespiti yöntemleri Türkçe metinler üzerinde ilk olarak bu çalışmada test edilmiştir.
- Farklı yöntemlerin birbiri ile kıyaslanmasında temel veri kümesi kullanılmıştır. Tezde literatürde yer alan çalışmaların aksine bilgi geri getirim sahasındakine benzer bir uygulama ile veri kümesi oluşturulmuştur. Bu yöntem, eşdizim tespit tekniklerinin tümüne kendilerine ait en iyi listeyi dâhil etme şansı sunması açısından bir ilktir.
- Tez kapsamında “Eşdizim Eğilimi Yöntemi” geliştirilmiştir. Bu yöntem uygulanması en kolay olmasına karşın başarımları düşük olan eşdizim tespit yöntemlerini dahi diğer yöntemler ile yarışır seviyede iyileştirmektedir.
- Önerilen “Eşdizim Eğilimi Yöntemi” dilden bağımsız bir yapıya sahiptir. Dolayısıyla farklı dillerde de uygulanması söz konusudur.

Tez anlatım akışı dört bölüme ayrılmıştır. İzleyen ilk bölümde metin analizi kavramı incelenmekte ve literatürde yer alan çeşitli çalışmalar sunulmaktadır. Bölüm 3’de işlev kelime –içerik kelime ayrımı tanıtılmakta ve önerilen yöntem tanıtılarak çalışma sonuçları verilmiştir. Bölüm 4, eşdizim tespiti çalışmasını içermektedir. Bu bölümde eşdizim kavramı sunulmakta, literatürde yer alan

yöntemler ve “Eşdizim Eğilimi Yöntemi” tanıtılmaktadır. Çeşitli metin ve derlemler için çalışma sonuçları verilmektedir. Tezin 5. ve son bölümünde ise sonuç ve çalışmayı ileriye götürebilecek öneriler yer almaktadır.

2. METİN ANALİZİ

2.1. Tanımlar

Dil, insanlar arasında anlaşmayı sağlayan tabii bir vasıta, kendisine özgü kuralları olan ve ancak bu kurallar çerçevesinde gelişen bir varlık, temeli bilinmeyen zamanlarda atılmış bir gizli anlaşmalar sistemidir (Ergin, 2000).

Dil kendisini sürekli yenileyen bir değerdir. Dil kendi içindeki bir kısım özellikleri sosyolojik, psikolojik vb. gibi pek çok unsurlara dayanarak geliştirirken bir kısmını öldürür, diğer bir deyişle evrimleşir (Aitchison, 1991; Bichakjian, 1988; Hawkins vd., 1992; Lass, 1980; McMahon, 1994; Ritt, 1995; Christiansen, 2003). Alman dilbilimci August Schleicher dili yaşam döngüsü olan organizmalara benzetmiştir. Schleicher, dili doğan, yetişkin hale gelebilmek ve karmaşık bir yapı oluşturmak için büyüyen, yaşlanan yani karmaşıklığını kaybeden ve ölen bir canlı olarak betimlemektedir (Ritt, 1995).

Fosket bilişsel psikoloji bakış açısıyla veri, bilgi ve enformasyon tanımlarını şöyle vermektedir:

“Veri doğrudan hiçbir aracı olmaksızın gözlemlenen her türlü gerçek veya gerçekler; Bilgi benim bildiğim şey, yani kişi olarak bildiğimiz şey veya şeyler; Enformasyon ise bizim bildiğimiz şeyler, yani paylaşılmış bilgidir.”

Fosket’in bilgi ve enformasyon tanımı dikkate alındığında dil bilginin enformasyona dönüşmesini sağlayan böylece insanlığın gelişmesine olanak tanıyan, milletlerin tarihi ve kültürel gelişimlerinin yansıması olan yegâne araçtır.

Bilginin dil vasıtasıyla aktarımı iki temel yöntemle gerçekleşmektedir: sözlü veya yazılı iletişim. Sözlü iletişim konuşma vasıtasıyla gerçekleşen dinamik bir bilgi aktarımıdır. Sözlü ifadeler, jestler ve tonlama gibi çeşitli fiziksel unsurlar ile tamamlanır. Bu sebeple sözlü ifadelerin taşıdığı anlam sadece sözcüklere (kelimelere) yüklenmemiştir (Atlıoğlu, 1994). Yazılı iletişim ise duygu ve

düşüncelerin yazılı semboller kullanılarak aktarılmasıdır. Yazılı iletişimde duygu ve düşünceler sözcükler ve çeşitli ünlem işaretleri ile ifade edilmektedir.

Bilgi iletişiminin sözlü veya yazılı yapılmasında, bir olguyu, düşünceyi, duyguyu açıklayan anlamlı bir bütünün oluşturulması gereklidir. Bir veya birden fazla kişi tarafından oluşturulan bu anlamlı bütün metin olarak adlandırılır. Metin en temel düzeyde harfler/sesler ve çeşitli sembollerden başlayarak, daha sonrasında kelimeler, cümleler ve paragraflar şeklinde çeşitli birimler ile yapılandırılmaktadır.

Metin, çok farklı ortamlarda dille iletişimde bulunmak amacıyla oluşturulan anlatma ve anlaşma aracıdır (Aktaş, 2009). Bu araç içinde yer alan cümleler anlam bütünlüğü taşımalarına rağmen çoğu zaman yalnız başlarına eksiksiz bir anlam oluşturmazlar. Yani herhangi bir konuda soruya gerek duyulmaması ve şüpheyi yer bırakılmamasını sağlamazlar. Metin cümleler ile örülmüş, dille anlatım ve anlaşmayı sağlayan, hedef kitle, amaç, konu, kullanılan iletişim kanalı gibi unsurlara göre değişen bir araçtır. Bu bağlamda cümleler dilbilim sınırlarında kalırken, metinler anlam bilim çerçevesinde değerlendirilmelidir (Aktaş, 2009). Aktaş (2009) çalışmasında bu fikrin metin kelimesinin sözlük anlamı içinde de yer aldığından bahsetmektedir

“Metin Arapçada sağlam ve dayanıklı anlamına gelen bir kökten türetilmiştir. Fransızcada ve diğer birçok batı dillerinde ise dokumak (örülmek) anlamına gelen “texte” kelimesiyle karşılanmaktadır. Türkçede kullanılan tekstil bu hususu açıkça ortaya koyar. Bunun için metin kelimesi dille gerçekleştirilmiş sağlam ve anlamlı örgüyü düşündürmektedir.”

Mantıksal, anlamsal ve dilbilgisel açıdan bağlantılı bir bütünlük olan metinde iki temel ilişki ağı mevcuttur: bağlaşıklık (İng. cohesion) ve tutarlılık/bağdaşıklık (İng. coherence) (Keçik ve Uzun, 2003; Adalar, 2005)

Metni oluşturan anlamlı birimler yani kelime ve cümleler metin içerisinde rastgele bir şekilde bir arada bulunmazlar. Metinlerde birimlerin belirli bir düzende dizilişini ve tematik bir bütünlük oluşmasını sağlayan bir takım ilişkiler, dilbilgisi kuralları vardır. İşte bu ilişkilere bağlaşıklık denir. Harris (1963)

çalışmasında cümleler arası bağlantıları etkileyen özel araçlara “bağlaç bağlar” (İng. coherent ties) adını vermiştir. Cümleler, deyimler kelimeler gibi anlam birimlerini bağlayan bu bağlar zarf, bağlaç, tanım edatı, zamirler gibi dilbilimsel araçlardır (Davies, 1997). Cümleler arası bağlılığı sağlamak için 8 tekniğin kullanılması önerilmektedir. Bu teknikler şu şekilde listelenmektedir:

1. **Kelime Tekrarlama:** Cümledeki bir kelimenin takip eden cümlede kullanılması
2. **Eşanlamlı Kelime Kullanımı:** Eğer aynı kelimenin tekrarı mümkün değilse eşanlamlı bir kelimenin tekrarlanması
3. **Zıt Anlamlı Kelime Kullanımı:** Cümledeki herhangi bir kelimenin zıt anlamlısının kullanılması
4. **Zamir Kullanımı:** Daha önce belirtilen bir kelime için zamir kullanılması
5. **İlişkili Kelime Kullanımı:** Bir arada gözlenme miktarları yüksek kelimelerin iki cümleyi bağlamak için kullanılması (örneğin doktor ve hasta kelimelerinin kullanılması)
6. **Numaralandırma:** Fikirler arası bağları oluşturmak için fikirlerin numaralandırılması
7. **Paralellik:** Bir cümle yapısının tekrarlanması
8. **Bağlaç Kelime Kullanımı:** Bağlaç sözcükler vasıtasıyla cümleler arası mantıksal ilişkiler sağlanması

Tutarlılık metni meydana getiren parçalar arasındaki anlam ilişkisine verilen addır. Tutarlılık her paragrafı cümleyi, terimi tek bir anlam bütününe katkı sağlayacak şekilde birbirine bağlayan pek çok etkenin bir ürünüdür. Van Dijk (1981) tutarlılığı oluşturan üç öge olduğunu belirtir. Bunlar “konu, fikir, sonuç veya söylemin özü” olarak listelenmiştir. Halliday ve Hasan’a (1976) göre ise tutarlılık kavramı anlamla ilgili bir kavram olup, bir söylem içerisindeki bir ögenin aynı söylem içerisindeki başka bir ögeye bağlı olduğu durumlarda belirmektedir. Özellikle yazılı iletişimde tutarlılığı sağlamak ve test etmek sözlü iletişime oranla daha güçtür. Çünkü sözlü iletişimde kullanılan pek çok ek araç, örneğin konuşmacıların iletmek istediği mesajın net olarak anlaşıldığını kontrol etmek için kullanacağı jestler gibi, yazılı anlatımda bulunmamaktadır.

Metni meydana getiren kendi içinde anlamlı bir birim olan parçalar arasındaki ilişki ağlarını belirlemek ve sorgulamak metnin varlık sebebi sayılabilecek unsurları ortaya çıkarır. (Aktaş, 2009). Filizok (2001), metnin anlamlı kılan bu öğeler arası ilişkilerin neticesinde anlamın oluştuğunu belirtmiştir. Metnin aktarmak istediği anlamın yani iletilmek istenen mesajın saptanması için bu birimler arasındaki ilişkiyi belirleme ve açıklama önemlidir. Bu ilişkiler ağını ortaya koymayı amaçlayan çalışmaların tümü metnin analizi kavramı içine dâhil edilir.

Metin analizi çalışmaları, metnin aktardığı enformasyonun tespiti için yürütülen her türlü çalışmayı içermektedir. Bu sebeple içerik analizi (content analysis) kavramı ile sık sık bir arada kullanılmaktadır. İçerik analizi metinlerden geçerli çıkarsamalar yapabilmek için bir grup işlemin kullanıldığı bir araştırma tekniğidir (Weber, 1990; Krippendorff, 2004). Analizin amacı metnin sıkıştırılarak daha az miktardaki içerik kategorilerine indirgenmesi ve bu şekilde değerlendirilmesidir (Stemler, 2001). Böylece metnin tümü yerine enformasyon içeriğini taşıyan yapılar değerlendirilerek, metin ile ilgili çeşitli soruların cevaplarının araştırılması hedeflenmektedir. İçerik analizinin cevap vermeyi amaçladığı soru grubu Lasswell tarafından şöyle ifade edilmiştir (Prasad, 2008).

“ KİM, NEYİ, KİME, NEDEN, NE ÖLÇÜDE söylüyor ve NASIL BİR ETKİ yaratıyor?”

İçerik analizinde nicel (qualitative) ve nitel (quantitative) olmak üzere iki farklı yöntem bir arada veya ayrı olarak kullanılabilir. Nicel analiz sistematik bir sınıflandırma süreci sonucunda metin içeriğinin subjektif olarak yorumlanmasıdır (Zhang ve Wildemuth, 2009). Çoğunlukla antropoloji, nicel sosyoloji ve psikoloji gibi alanlarda fiziksel mesajların altında yatan anlamın yakalanmaya çalışılmasıdır. Nitel analizde ise mesajların belirlenmiş karakteristiklerini objektif ve sistematik bir şekilde tespit ederek çeşitli çıkarsamalar yapılır. Nitel ve nicel analizin temel farkları Zhang ve Wildemuth çalışmasında (2009) şu şekilde listelenmiştir.

- Nitel analiz tümdengelim, nicel analiz ise tümevarım prensibini kabul eder.

- Nitel analizde tesadüfî örnekleme (random sampling) yöntemi veya başka bir istatistiksel yöntemle veri kümesi oluşturulur. Nicel analizde ise bilinçli olarak araştırma soruları hakkında bilgi verebilecek metinler seçilir.
- Nitel analiz sonucunda ürün olarak tanım ve tiyolojiler¹ verilir. Bu ürünler vasıtasıyla analiz edilen metnin/metinlerin yazarı/yazarlarına ait bakış açısının araştırmacı kadar analiz sonuçlarını okuyan kişilerce de anlaşılması sağlanır. Nicel analizde ise istatistiksel bir takım teknikler kullanılarak sayılar üretilir. Bu yöntemde sonuçların istatistiksel olarak geçerliliği ve genellenebilir olması beklenir.

2.2. İlk Çalışmalar

Otomatik içerik analizi yöntemi temelde metinlerdeki bağlama ait anahtar kelimelerin (İng. Key words in context-KWIC) otomatik olarak saptanması ve metinlerin bu kelimeler ile ifade edilmesidir (Van Rijsbergen, 1979; Stemler, 2001). Metinlerin içerdiği anlamlı birimler ile ifade edilerek değerlendirilmesi aslen ilk olarak Zipf tarafından ortaya koyulmuştur (1949).

Zipf “Human Behavior and Principle of Least Effort” (1949) kitabında, en az gayret prensibini esas alarak dilin farklı özelliklerini modellemiştir. Bu yaklaşımda kelimeler hedeflere ulaşmak yani belli bir anlamı aktarmak için kullanılan araçlardır. Konuşmacı bu aracı kendi açısından en ekonomik şekilde kullanarak dilediğini iletmek, dinleyici ise yine aracı en ekonomik şekilde kullanarak anlatılanı anlamak eğilimindedir. Konuşmacı açısından ekonomi, bir kelimeye birçok anlam yüklenerek sağlanabilir. Dinleyici ise algılama çabasını azaltmak amacıyla her farklı anlam için farklı bir kelime duyma isteğinde olacaktır. Bu iki ters kuvvetin çarpışması ile dildeki denge oluşmaktadır. Zipf, bu kuvvetlerin oluşturduğu denge durumunun dilin farklı özelliklerini belirlediğini ve metinlerin içerik bilgisinin aktarılmasında önemli olduğunu ileri sürmüş ve dilin birbiriyle ilişkili farklı özelliklerini 4 kanun altında incelemiştir.

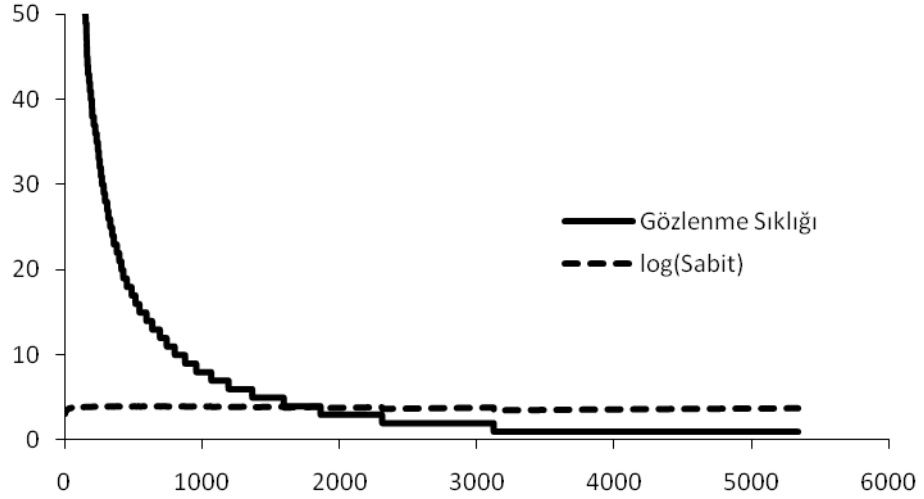
¹ Aynı katman içinde birbiriyle ilişkili olduğu saptanan buluntuların biçimsel özelliklerine göre sınıflandırılması

Zipf'in ilk çalışmasında derlem içindeki her kelime sayılarak gözlenme sıklık değerleri azalacak şekilde dizilip, her bir kelimeye bu dizideki yerine bağlı olarak bir sıra numarası verilmektedir. Böylece yüksek sıklıkla gözlenen kelimeler düşük sıra numaralarına, düşük sıklıkla gözlenen kelimeler yüksek sıra numaralarına sahip olmaktadır. Böyle bir numaralandırma düzeni içinde bir metinde veya çeşitli metinler içeren bir derlemde yer alan son (en büyük) sıra numarası metnin/derlemin kelime dağarcık miktarını vermektedir.

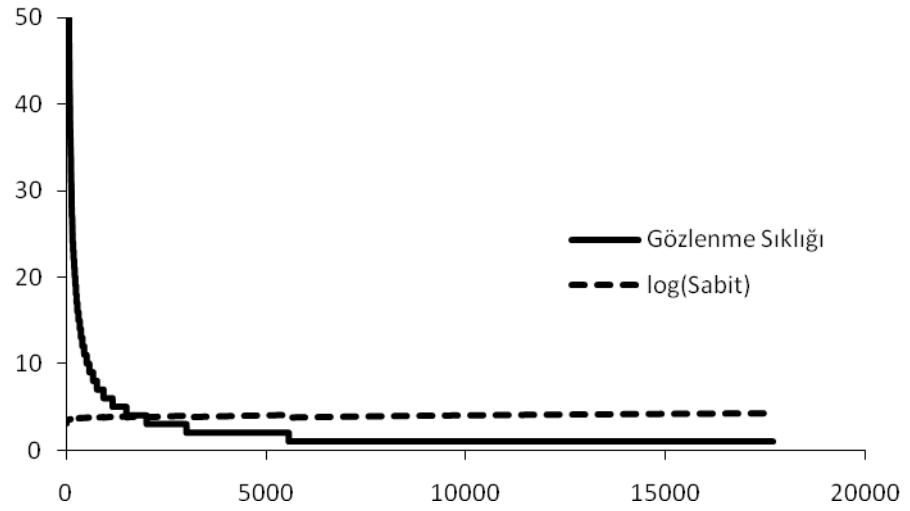
En az gayret ilkesine göre dinleyici, kelime dağarcığı miktarını yani son sıra numarası değerini arttırmak amacındadır. Konuşmacı ise kelimelere ait sıklık değerlerini arttırarak çok anlamı tek bir kelimeye yüklemek eğilimindedir. Bu güç savaşı sonucunda kuvvetlerin dengesi sağlanmaktadır. Bu denge sayesinde bir kelimenin gözlenme sıklık değeri (frekans) ile sıra numarası arasında şu ilişki oluşmaktadır (Zipf 1. Kanun):

$$f \times r \cong C \quad (2.1)$$

Denklemden f gözlenme sıklığı değerini, r sıra numarasını C ise sabit bir değeri vermektedir. Şekil 2.1 ve Şekil 2.2'de özellikleri ileriki bölümlerde açıklanacak (Bölüm 3.4.3) OSTAD derleminin sırasıyla gövde ve yüzey formu için sıra numarası-gözlenme sıklığı grafikleri verilmiştir. Yatay eksen sıra numarası (r) bilgisini taşımaktadır. Grafiklerde C sabitinin her bir kelime için logaritma değerini gösteren eğri de bulunmaktadır.



Şekil 2.1 OSTAD derleminin gövde formu için gözlenme sıklığı – sıra numarası grafiği



Şekil 2.2 OSTAD derleminin yüzeysel formu için gözlenme sıklığı – sıra numarası grafiği

Zipf 1. Kanunu üzerinde yapılan çalışmalar ve iyileştirmelerde Zipf'in tanımladığı ilişki pek çok farklı formda ele alınmıştır (Mandelbrot 1952, 1959; Herdan 1960; Samuelson 1996; Powers 1998; Baayen, 1996; Kornai 2002). Bu düzeltme ve iyileştirme çalışmaları sonucunda geliştirilen denklemlerden en yaygın kabul göreni uç frekans değerleri için daha başarılı sonuçlar üreten Mandelbrot denklemidir:

$$\log(f_r) = \log(P) - B \log(r + W) \quad (2.2)$$

Mandelbrot denkleminde, f_r frekans, $\log(P)$ bir sabit, W uç frekans değerlerindeki düzeltmeleri sağlayan derlem bağımlı bir değerdir.

Zipf 2. Kanunu kelime dağarcığının matematiksel olarak hesaplanmasını amaçlar. Kanun kısaca şu şekilde tanımlanır:

N büyüklüğünde bir derlemde, i kere gözlenen farklı kelime sayısı $(V(i, N))$, gözlenme sıklık değeri (i) ile ters orantılıdır

İlişkinin matematiksel ifadesi denklem (2.3)'de verilmektedir.

$$\log(i) = K_N - D_N \log(V(i, N)) \quad (2.3)$$

K_N ve D_N , derlem büyüklüğüne (N) bağlı olan sayısal sabitlerdir. (2.4) denklemi ile verilen ilişki sonucunda bir derleme ait kelime dağarcığı ise tüm i değerleri için $V(i, N)$ değerinin toplanması ile hesaplanabilir.

$$V = \sum_i^n V(i, N). \quad (2.4)$$

Zipf'in "Anlamlar Kanunu" olarak da adlandırılan 3. kanunu kelime sıklıkları ve bu kelimelerin taşıdıkları farklı anlamların miktarı arasındaki ilişkiyi gösterir. Bu ilişkinin belirlenmesi çalışmasında, en az gayret ilkesi gereği çarpışan zıt kuvvetlerin etkisi en sık rastlanan kelime üzerinde değerlendirilmiştir. En sık rastlanan kelimenin frekans değerine F_1 , taşıdığı farklı anlam sayısına ise m_1 denilirse şu ilişki ortaya çıkar:

$$F_1 = m_1 \times f_1 \quad (2.5)$$

Burada f_1 , m_1 adet anlamın ortalama gözlenme miktarını yani frekans değerini ifade eder. Konuşmacı tüm anlamları bir kelimeye yüklemek amacıyla m_1 değerini arttırmak isteyecektir. Dinleyici ise bir kelimeye düşen anlam miktarını azaltmak amacıyla f_1 değerini arttırmak isteyecektir. Deneysel çalışmalar sonucu kuvvetlerin etkisiyle bu iki değişken arasında hiperbolik bir

ilişki olduğu ortaya çıkmıştır. Sonuçta m_1 değeri f_1 değerine yaklaşarak denklem (2.6) ve (2.7)'de verilen ilişkileri doğurmaktadır.

$$F_1 = m_1 \times m_1 \quad (2.6)$$

$$\sqrt{F_1} = m_1 \quad (2.7)$$

En yüksek sıklıkla rastlanan kelime için verilen bu ilişkiler diğer sıklıklardaki kelimeler için genelleştirilerek “kelime anlam miktarı (m), sıklık-frekans (f) değeri ile doğru orantılı, sıra numarası (r) ile ters orantılı şekilde artar” şeklinde ifade edilebilir. Matematiksel olarak ise bu ilişki şöyle tanımlanır:

$$m \propto \sqrt{f} \quad (2.8)$$

$$m \propto 1/\sqrt{r} \quad (2.9)$$

Zipf’in 4. kanunu “Öbekleşme Kanunu” olarak da tanınır. Konuşmacı veya yazarın bir konuyu aktarırken kullandığı bir kelimeyi tekrar kullanarak sarf ettiği gayreti azaltma eğiliminde olacağı prensibi bu kanunda esas alınmıştır. Bu prensibin yazılı metinlerdeki etkisi şu şekilde özetlenebilir:

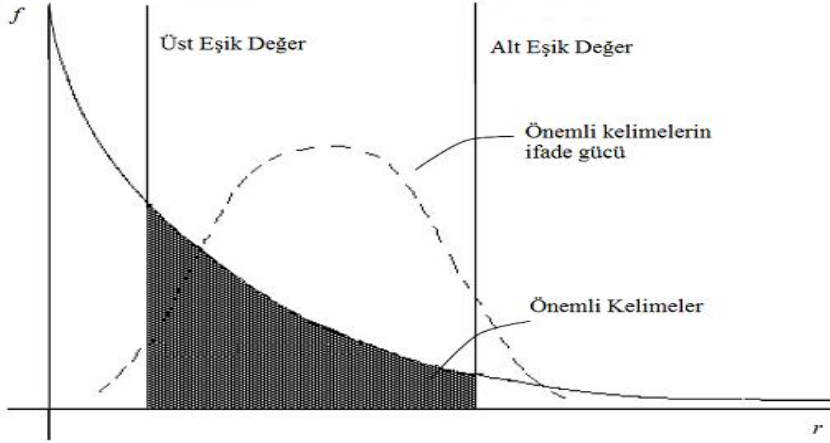
Bir içerik kelime, metinde bir kere yer almışsa, aynı konunun devam ettiği cümlelerde, paragraf veya sayfalarda yüksek olasılıkla tekrar gözlenir.

Öbekleşme Kanunu, kelimelerin tekrar gözlenme aralığını ölçme ve bir içerik kelimenin rastlanma desenini belirlemeyi sağlamaktadır. Kanuna göre kelimenin gözlendiği yerler arası aralık uzunluğu (I_f) ile bu uzunluğa sahip aralık miktarı (N) arasında denklem 2.10’da ifade edildiği gibi ters bir orantı vardır.

$$N^{-p} \propto I_f \quad (2.10)$$

Zipf’in çalışmasından yola çıkarak metinlerin temsilini sağlayan KWIC (Key Word in Context) kavramı ilk kez otomatik metin analizinin temellerini

attığı kabul edilen Luhn tarafından kullanılmıştır (Van Rijsbergen, 1979). Luhn çalışmasında bir kelimenin öneminin içinde bulunduğu metindeki gözlenme miktarı ile ilgili olduğunu belirtmiştir. Zipf'in dikkat çektiği gözlenme sıklığı-sıra numarası eğrilerinde çok ve az sıklıkla gözlenen kelimelerin önemsiz olduğunu belirtmiştir. Şekil 2.3'de şematik olarak bir gözlenme sıklığı-sıra numarası ve Luhn'un kelimelerin önemini ifade eden önem eğrisi verilmiştir.



Şekil 2.3 Kelime gözlenme sıklık-sıra numarası eğrilerinde önemli kelimelerin ifade gücünün şematik gösterimi

Şekil 2.3'de dikkat çeken husus bir üst ve alt sıklık eşik değeri arasında kalan kelimelerin önem arz ettiği, diğer kelimelerin ise metni ifade etmekte güçsüz olduklarının gösterilmiş olmasıdır. Luhn bu eşik değerler arasında kalan kelimelerin KWIC olduğunu vurgulamış ve metindeki anlamı taşıdıklarını belirtmiştir. İki eşik sıklık değeri arasında kalan kısımda şekilden de görüldüğü üzere kelimelerin ifade gücü belirli bir noktada en yüksek değerine ulaşır sonra azalmaktadır. Metin analizindeki temel amaç bu ifade gücü yüksek kelimelerin yakalanmasıdır. Grafiğin üst eşik sıklık değerine kadar olan kısmında metinlerde sıklıkla rastlanan ve sadece harç görevi yapan işlev kelimeler, alt eşik değerinden sonra ise metin içinde az sıklıkla gözlenen önemsiz kelimeler yer almaktadır.

Luhn'un ortaya attığı fikir, yani kelimelerin gözlenme sıklığı ile bağlantılı olarak anlam taşıdıkları fikri pek çok bilim alanında kullanılmış ve geliştirilmiştir. Sosyal bilimler, iletişim, psikoloji, siyasal bilimler, tarih, dil bilim, hesaplamalı dil bilim, bilgi geri getirim gibi farklı disiplinlerde bu temel fikirden yola çıkarak geliştirilen metin/içerik analizi çalışmalarına rastlanmaktadır.

Sosyal bilimler alanında ayrımcılık, kültürel sembollerdeki değişim, sosyal eğilimler ve değişimlerin incelenmesi gibi çeşitli konularda metin/içerik analizi çalışmaları yapılmıştır. Örneğin kültürel sembollerdeki değişimin tespiti için Taviss (1969) 1900-1950 yılları arasında yayınlanan popüler romanları inceleyerek Amerikan halkının orta sınıfında sosyal yabancılaşmanın azalarak bireysel yabancılaşma olgusunun arttığını ortaya koymuştur (Prasad, 2008). Benzer şekilde Lowenthall (1944) popüler dergilerdeki kahraman tanımlarındaki değişimleri incelemiş; analiz sonucunda işadamları ve profesyonellerin yerini eğlence sektöründe çalışan insanların aldığını göstermiştir (Prasad, 2008).

Metinlerin anlaşılması, işlenmesi, geliştirilmesi ve otomatik olarak üretilmesi gibi çeşitli çalışmaları bünyesinde barındıran hesaplamalı dil bilim ve bu çalışmaların derlemler üzerinden yürütüldüğü derlem dil bilim alanında metin analizi sıklıkla temel bir işlem basamağı olarak karşımıza çıkmaktadır. Örneğin metin yazarının tespiti çalışmalarında aday yazarların eski yazılarından elde edilen anlam taşıması muhtemel öğelere (kelime veya cümle gibi) ait çeşitli istatistikler kullanılabilmektedir. Bunlara örnek olarak Diri ve Amasyalı (2003), Putnins vd. (2005), Martindale ve McKenzie (1995) gibi pek çok çalışma verilebilir.

Bilgi geri getirim (BGG) Manning vd. (2008) tarafından

“Büyük koleksiyonlardan enformasyon ihtiyacını karşılayacak tanımlı bir yapıya sahip olmayan (genellikle metin halinde) materyalin (genellikle belge) bulunması”

şeklinde tanımlanmıştır. Bilgi geri getirim sistemlerinde tanımlı yapıya sahip olmayan materyal yani metinlerin muhteviyatının tespiti gereklidir. Bu işleme indeksleme adı verilir. ANSI 1968 standardında, indeksleme için verilen tanım şöyledir (Dinçer, 2004):

“İndeksleme: Bilgi kayıtlarından enformasyon muhteviyatını analiz etme ve indeksleme sisteminin dili ile bu enformasyon muhteviyatını beyan etme işlemidir. Bu işlem sırasıyla: bir belgede indekslenebilir muhteviyatın seçilmesi ve indeksleme sisteminin dili ile bu muhteviyatın beyanının sıralı bir liste haline getirilmesi adımlarından oluşur.”

İndeksleme sisteminin bir çıktısı olarak ortaya çıkan, bir metnin içeriği hakkında kayda değer bir enformasyon miktarını içeren ifade indeks terim olarak adlandırılır (Lahtinen, 2000). Metin analizi metinlerde anlam taşıyan öğelerin belirlenmesini amaç edinmesi sebebiyle bilgi geri getirim disiplini içinde indeks terimlerin tespiti açısından önemlidir. Örneğin metin analizinde ilk çalışma olarak görülen Luhn'un anlam-gözlenme sıklığı ilişkisi dahi basit görünmesine rağmen metnin taşıdığı anlamı ifade eden kelimelerin sıklık değerlerine bağlı olarak saptanabilmesini sağlaması sebebiyle bilgi geri getirim alanında basit bir indeksleme imkânı sunmaktadır.

3. İŞLEV KELİMELERİN TESPİTİ

3.1. İşlev Kelime Tespiti Problemi

Yazılı metinlerde anlam bütünlüğü taşıyan en küçük parçalar cümlelerdir. Her bir cümle bir konu, bir olay hakkında bilgi veren, bir kavram açıklayan kelime (sözcük), simge ve simge birliklerinden oluşur. Cümle içinde her bir kelimenin farklı görevleri vardır ve dil içinde rastlanan her bir kelime çoğu zaman farklı cümlelerde farklı görevler alır. Cümlelerin, paragrafların veya metinlerin anlam bütünlüğünü, çeşitliliğini sağlayan da içeriğindeki kelimelerin anlam miktarına katkısının değişkenlik gösterebilmesi durumudur.

Bir cümlede gözlenen her kelimenin cümlelerin genel anlamına veya içinde bulunduğu metne kattığı enformasyon miktarı aynı değildir. Dil içindeki dilbilgisi yapısının korunmasını ve akıcılığı sağlayan, anlama az veya hiç katkıda bulunmayan kelimeler vardır. İşlev kelime (fonksiyon kelime, durma kelimesi) olarak adlandırılan bu kelimeler az miktarda sözcüksel anlama veya belirsiz anlama sahip olan kelimelerdir (Manning and Schütze, 1999). İşlev kelimeler, cümlelerin dilin kurallarına göre oluşmasını sağlar ve çoğu zaman sadece harç görevi yaparlar. Bu kelime grubu az miktarda kelime içerse de metinlerin büyük bir kısmını işgal ederler. Kornai (2000) çalışmasında işlev kelimelerin bir metnin yaklaşık yarısını oluşturduğu üzerinde durmuştur. Diğer kelime grubu ise başlı başına anlam taşıyan ve dil içinde yaratılan her metnin taşıdığı enformasyon miktarını arttıran bir niteliğe sahiptir. Bu grupta yer alan kelimeler “*içerik kelime*” (İng. content word, content bearing word) olarak adlandırılır. İçerik kelimeler genelde bir kavram veya olayı simgeleyen isim veya fiillerden oluşur.

İşlev kelimeler dil içinde genelde zarf, edat, bağlaç ve zamir olarak kullandığımız kelimelerdir. Türkçe’de “*ve, veya, ile, ki, ben, sen, o, gibi*”, İngilizce’de ise yaygınlıkla kullanılan işlev kelimelere örnek olarak “*and, a, be, but, the, it, in*” kelimeleri verilebilir.

İşlev kelimeler dil bilim ve özellikle doğal dil işleme (İng. Natural Language Processing) çalışmalarında önemli bir yer tutmaktadır. Bu kelimelerin dilin genel kullanımı dikkate alınarak seçilmesi ve listelenmesi araştırmaların verimliliğini doğrudan etkilemektedir. İşlev kelimeler çoğu zaman yüksek miktarlarda kullanıldıkları için metin içinde çok yer işgal etmektedirler. Bu durum, özellikle amacı en uygun materyali bularak ilgisiz olanların ayıklanması olan bilgi geri getirim sistemlerinde önemli bir zaman ve kaynak sarfiyatına sebep olur (Dolamic and Savoy, 2009). Bilgi geri getirim dışında makine tercümesi, otomatik özetleme vb. alanlarda da işlev kelimelerin ayıklanması temel bir işlem basamağı olarak karşımıza çıkmaktadır (Bond and Shirai, 2003; Martinez-Santiago et al., 2004; Jiang et al., 2009; Radev and Fan, 2000; Argamon and Levitan, 2005 ve diğerleri). Bu ayıklama işlemi kimi zaman işlev kelime listesi kimi zaman ise kelime türü bilgisi (Witten et al., 1999) kullanılarak yapılmaktadır.

Dil içindeki işlev kelime, içerik kelime ayrımı günümüze kadar yürütülen çalışmalarda görünme sıklığına bağımlı olarak yapılmıştır. Herdan (1960) ile başlayarak içerik ve işlev kelimelerin farklı iki grup içinde değerlendirilmesi söz konusu olmuştur. Bu iki grup görünme sıklığına bağlı olarak belirlenmektedir. Görünme sıklığı yüksek olan kelimeler işlev, görünme sıklığı düşük olan kelimeler içerik kelime olarak adlandırılmaktadır. Ancak yapılan çalışmalar sonucunda sadece sıklık bilgisinin yeterli olmadığı yani yüksek sıklıkla gözlenen bazı kelimelerin aslında işlev kelime olmadıkları gözlemlenmiştir.

İngilizce'ye oranla üzerinde daha az çalışma yapılan Türkçe'de işlev kelimelerin tespiti için yüksek sıklıkla rastlanan belli sayıda veya oranda kelimenin işlev kelime olarak kabulü (Can et al., 2008; Kumova Metin 2008); yüksek sıklıkla rastlanan kelimeler içinden hatalı olanların ayıklanması ile işlev kelime listesinin oluşturulması (Can et al., 2008); bağlaç, edat ve benzeri gibi işlev kelime olması muhtemel kelimelerin sözlükten saptanarak bir liste oluşturulması (Alpkoçak ve Özkarahan, 1995) gibi yöntemlere rastlanmaktadır. Yöntemlerin pek çoğu ya otomatik değildir ya da yarı otomatiktir ve hataların elle ayıklanması gerekliliği söz konusudur. Ayrıca yöntemlerin çoğunda sabit bir liste tüm derlem ve metin grupları için kullanılmaktadır.

Tez kapsamında işlev-içerik kelime gruplarının ayıklanmasında kelime sıklık bilgisine ek olarak kelimelerin derlem veya metin içinde taşıdıkları bir takım özelliklerin de etkili olduğu kabulü ile bir grup özelliğın bir arada değerlendirildiğı bir yöntem sunulmuştur. Doğrusal bir denklem ile özelliklerin bir arada temsil edilebildiğı diskriminant analizi yöntemi önerilmektedir. Yöntem Türkçe bir derlem üzerine uygulanmış ve sonuçlar pek çok çalışmada sıklıkla kullanılan kelime sıklık ve tf-idf yöntemleri ile kıyaslamalı olarak sunulmuştur.

İzleyen bölümlerde işlev kelime tespiti literatürü, diskriminant analizinin matematiksel temelleri, analizin işlev kelime tespitinde kullanımı ve sonuçlar bulunmaktadır.

3.2. Literatür

Geleneksel olarak işlev kelime listeleri metin veya derlemlerde sıklıkla rastlanan kelimelerden oluşurlar. İşlev kelimelerin tespitinde kelimelerin gözlenme miktarına bağımlı olarak geliştirilmiş ve yaygın olarak kullanılan iki yöntem bulunmaktadır: kelime sıklık yöntemi (tf) ve tf-idf (kelime sıklığı- belge sıklığının tersi).

Kelime sıklık yöntemi, kelime gözlenme değerine göre azalan sırada listelenmiş kelimelerden yüksek sıklık değerine sahip olanların işlev olarak atanmasını öngörür. Yöntemin uygulanması esnasında derlemdeki her kelime için sıklık değeri belirlenerek kelimeler sıklık değeri azalacak şekilde listelenir, yüksek sıklıkla rastlanan kelimelerin işlev kelime olduğu kabul edilir. Yöntemin iki temel mahzuru vardır. İlki, bir metin içinde anlamı pekiştirmek amacıyla sıklıkla kullanılmış ve esasen içerik kelime olan kelimelerin sıklık değerlerinin yüksek olması sebebiyle işlev olarak atanmasıdır (Manning and Schütze, 1999). Diğeri ise sıklık değerine göre azalan sırada kelimeler içeren listede bir eşik değere bağılı olarak işlev – içerik ataması yapılması durumudur. Bu eşik değeri işlev ve içerik kelime gruplarını birbirinden kesin çizgilerle ayıran ve yöntemin başarısını doğrudan etkileyen bir unsurdur. Eşik değeri, belirli bir sıklık değeri olarak kabul edilebilir veya aday listeleri içinde belirli bir eşik sıra değerinden düşük sıra değerine sahip adaylar işlev kelime olarak etiketlenebilir. Eşik değerin

belirli bir sıklık değeri olduğu durumda eşik değerin düşük seçilmesi çok miktarda işlev kelime ataması yapılmasını dolayısıyla derlem içinde yer alan işlev kelimelerin birçoğu yakalanırken hatalı olarak işlev kabul edilen içerik kelime miktarının da artmasına sebep olur. Aksi durumda yani eşik değerin yüksek bir değer olduğu durumda ise az sayıda işlev kelime yakalanacak ve işlev kelimeler hatalı olarak içerik kelime kabul edileceklerdir. Eşik değerin sabit bir sıra numarası olarak kabul edildiği durumda da benzer sorunlar söz konusudur. Ancak bu mahsurlarına rağmen kelime sıklık yöntemi pek çok farklı dilde yapılan çalışmalarda halen kullanılmaktadır. (Lazarinis, 2007; Savoy, 1999; Can et al., 2008; Zou et al., 2006). Bu çalışmaların bir kısmı gövdeleme, kelime türü belirleme, sıklık değerinin normalize edilmesi gibi çeşitli yöntemlerle iyileştirilmeye çalışılmıştır.

Bilgi geri getirim ve metin madenciliği gibi çalışmalarda işlev kelime tespiti için kullanılan yöntem ise tf-idf yöntemidir. Bu yöntemde birden fazla metin gereklidir. Yöntem bir kelimenin bir metin grubu veya derlem içinde ne denli önemli olduğu konusunda bir ağırlıklandırma sunmaktadır. T_j terimi için tf- idf ağırlığı (w_{ij}) şu şekilde hesaplanır

$$w_{ij} = tf_{ij} * \log_2\left(\frac{N}{n}\right) \quad (3.1)$$

tf_{ij} , T_j teriminin metindeki gözlenme miktarını yani sıklık değerini, N derlem veya metin grubundaki toplam metin sayısını, n ise T_j teriminin en az bir kere gözlendiği farklı metin sayısını ifade etmektedir. Derlemde yer alan her bir kelime/terim için tf-idf değeri hesaplanarak, her metin için bu değerler artan sırada listelenir. tf-idf değeri düşük olan kelimeler işlev, yüksek olanlar ise içerik kelime olarak kabul edilir. Tf- idf yöntemin kelime sıklık yöntemine oranla daha yüksek başarımlı değerleri üretmektedir. Çünkü bir metinde anlamı pekiştirmek amacıyla sıklıkla gözüken içerik kelimeler diğer metinlerde sıklıkla gözükmeyeceklerdir böylece idf değerleri düşük olacak ve listede alt sıralarda kalacaklardır.

Kelime sıklık yöntemi ile kıyaslandığında daha başarılı olarak kabul edilse dahi tf-idf yönteminin de birtakım zayıf noktaları mevcuttur. Örneğin yöntem belge yani metin sınırları belirlenmemiş bir metin grubunda uygulanamaz. Kelime sıklık yöntemine benzer şekilde sıralı liste içinde işlev ve içerik kelimeleri iki gruba ayrılması için bir eşik tf-idf değeri belirlenmesi zorunludur. Bunlara ek olarak, üzerinde çalışılan metin grubu veya derlem benzer konularda metinleri içermekte ise bir takım içerik kelimeler metinlerin büyük bir kısmında gözleneceği için işlev olarak etiketlenecektir. Örneğin sağlık ile ilgili metinleri içeren bir derlemden “doktor”, “hemşire” gibi içerik kelimeler işlev olarak işaretlenecektir.

Literatürde sıklıkla rastladığımız diğer bir işlev kelime tespiti yöntemi ise önceden hazırlanmış sabit işlev kelime listelerinin kullanılmasıdır. Bu listeler oluşturulurken kelimeler bir sözlük veya derlemden çekilebilir veya araştırmacı kendi bilgisi dâhilindeki işlev kelimeleri listeye ekleyebilir. Bu yöntemdeki sorun işlev kelime listesinin sözlük/derlem veya araştırmacının kelime dağarcığı ile sınırlı olması durumudur. Bu yöntemde alan bağımlı işlev kelimelerin tespiti söz konusu olamaz. Örneğin sadece web belgeleri üzerinde gerçekleşen bir indeksleme çalışması için “web”, “e-posta” gibi kelimeler işlev olarak kabul edilmelidir. Sabit liste kullanılan yöntemlerde bu gibi kelimelerin liste içine dâhil edilmesi söz konusu olmaz.

İngilizce metinlerde işlev kelimelerin otomatik tespiti Tsz-Wai Lo, He ve Ouins (2005) çalışmasına tartışılmıştır. Bu çalışmada önerilen yöntem, terim tabanlı rassal örnekleme (İng. term based random sampling), geliştirilmiş bir gözlenme sıklığı ölçüsüne dayanmaktadır. Blanc ve Dominey (2004) ise daha bilişsel temelli bir yöntem önermişlerdir. Araştırmacılar, vurgu, tonlama ve benzeri gibi bürünsel (İng. prosodic) özelliklerin işlev ve içerik kelime ayrımında etken olduklarını savunmaktadırlar. Diğer yöntemlerden farklı olarak bu yöntemde ilgili özelliklerin temin edilebilmesi için yazılı metinler yerine konuşma derlemlerine ihtiyaç duyulmaktadır. Hao ve Hao (2008) ise çalışmalarında ki-kare istatistiklerine dayalı bir yöntemle Çince için işlev kelime listesi oluşturmuşlardır.

Günümüzde bilgi geri getirim çalışmalarının büyük bir kısmı web belgeleri üzerinde yürütülmekte olduğu için bu tip metinler için işlev kelimelerin tespit

edilmesi zorunlu hale gelmektedir. Sinka ve Corne (2003) çalışmasında entropi, tabanlı bir yöntem ile web belgelerindeki işlev kelimelerin tespit edilmesi söz konusudur. Bu işlev kelime listelerinde “e-posta” gibi işlev kelimeler bulunmaktadır.

Farklı dillerdeki metinler incelendiğinde işlev kelimelerin sadece gözlenme sıklığı özelliği ile saptanamayacağı görülmektedir. Bu sebeple tez kapsamında kelimeye ait olan farklı bir takım özelliklerin bir arada kullanılabileceği diskriminant analizine dayanan bir yöntem sunulmuştur. Bu yöntem farklı özellikleri doğrusal bir denklemde birleştirerek bir kelimenin işlevlik kararını vermektedir.

Diskriminant analizi kullanılarak gerçekleştirilmiş benzer bir çalışma Dillon ve Federhart’ın araştırmasıdır (1982). Bu çalışmada kelimelerin farklı belgelerdeki dağılımları dikkate alınarak indeks kelimelerin tespit edilmesi söz konusudur.

3.3. Diskriminant Analizinin Temelleri

Diskriminant analizi, kategorik bağımlı değişkenler ile metrik bağımsız değişkenler arasındaki ilişkileri tahmin etmeyi amaçlayan çok değişkenli istatistik tekniklerinden biridir. Analiz normal dağılım, gruplar arası homojen varyans ve kovaryans, bağımsız değişkenler arasında bağımsızlık kriteri gibi varsayımlar içermektedir. Diskriminant analizinin temel amaçları şöyle listelenebilir (Kalaycı vd., 2006)

- Grup üyeliğini tahmin etmek yani bir verinin hangi değişken grubuna gireceğine karar vermek
- Verilerin diskriminant fonksiyon eşitliği kullanılarak gruplara ayrılması (sınıflandırma)
- Bağımsız değişkenlerin aritmetik ortalamalarının gruplar arasında nasıl değiştiğini tespit etmek

- Bağımlı değişkenin varyansının ne kadarının bağımsız değişkenler tarafından açıklanabileceğini belirlemek
- Grupları ayırmada etkili olan ve olmayan değişkenleri belirlemek
- Verilerin tahmin edilen gibi sınıflanıp sınıflanmadığını test etmek

Analiz doğrusal veya kuadratik olarak iki farklı şekilde gerçekleştirilebilmektedir. Bu çalışmada doğrusal diskriminant analizi işlev kelimelerin tespiti ve kelimelerin işlev –içerik olarak gruplanmasını sağlayan bağımsız değişkenlerin değerlendirilmesi amacıyla uygulanmıştır.

Doğrusal analiz çok sayıda grubun üyelerini birbirinden ayırmak için farklı özelliklerine (x_i) ağırlıklı katsayılar atayarak birden fazla özelliğe ait değerlerin tek bir değer olarak ifade edilmesini sağlamaktadır. Bu ağırlıklı katsayılar grupları birbirinden ayırmada en etkin olan değerler olarak belirlenmektedir (Overall and Klett, 1972). Doğrusal diskriminant analizi sonucunda p farklı özellik değeri verildiğinde aşağıdaki gibi bir diskriminant fonksiyonu elde edilmektedir.

$$y = a_1x_1 + a_2x_2 + \dots + a_px_p + c \quad (3.2)$$

burada $a_1, a_2, \dots, a_p$ terimi her bir gözlemden elde edilen p orjinal değere uygulanacak ağırlıklandırma katsayılarını, c sabit değeri ifade etmektedir. Analizin amacı grup içi varyansı minimize ederken gruplar arası varyansı maksimize edecek optimum $a_1, a_2, \dots, a_p$ katsayıları üretmektir. Yani gruplar arası F oranının veya t istatistiğinin maksimize edilmesi istenmektedir. Maksimize edilmesi istenen fonksiyon ilk olarak R. A. Fisher tarafından tanımlanmıştır, bu fonksiyon gruplar arası varyans ile grup içi varyansın oranıdır.

$$f(a_1, a_2, \dots, a_p) = \frac{n_1n_2}{n_1 + n_2} \frac{(a_1d_1 + a_2d_2 + \dots + a_pd_p)^2}{\sum \sum c_{ij}a_ia_j} \quad (3.3)$$

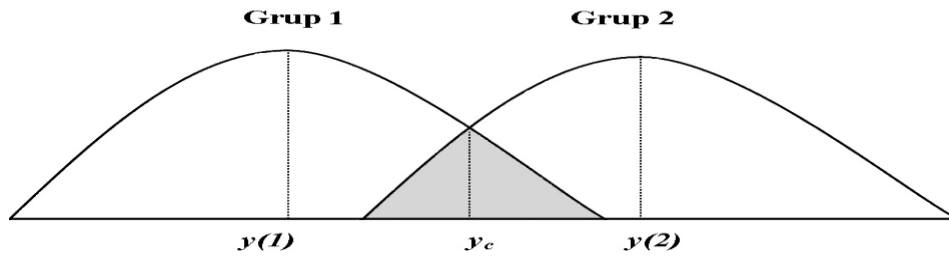
Bu kriter fonksiyonunda $[d_1d_2, \dots, d_p]$ vektörü p adet orijinal ölçüm değerinin ortalama farklarına ait vektördür, c_{ij} değerleri ise grup içi kovaryans

matrisini oluşturmaktadır. n_1 ve n_2 değerleri ise sırasıyla bir ve ikinci gruba ait örnek miktarını simgelemektedir. Fonksiyonun payda kısmında yer alan $\sum \sum c_{ij} a_i a_j$ terimi, a_i 'lerin ağırlıklandırma katsayıları olduğu doğrusal bir kombinasyonun grup içi varyansını simgelemektedir. Kriter fonksiyonunun pay kısmı ise gruplar arası varyans ile doğru orantılıdır. Buradaki problem çok sayıdaki veri için ortalama vektörü ve grup içi varyans-kovaryans matrisi verilmişken kriter fonksiyonunu maksimize edecek katsayıların saptanmasıdır. Bu problemin çözümü için matris formunda verilen şu eşitliğin çözülmesi gerekmektedir (Overall and Klett, 1972):

$$\begin{bmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{12} & c_{22} & \dots & c_{2p} \\ \dots & \dots & \dots & \dots \\ c_{p1} & c_{p2} & \dots & c_{pp} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \dots \\ d_p \end{bmatrix} \quad (3.4)$$

ve matris olarak $C\mathbf{a}=\mathbf{d}$

Çözüm için grup içi varyans-kovaryans matrisinin tersi alınarak eşitliğin her iki tarafı ile çarpılır ve $\mathbf{a}$ değerleri saptanır. Sonuçta iki grubun söz konusu olduğu durumda Şekil 3.1'de görüldüğü gibi iki farklı bölgede yoğunlaşmış olan her bir grup için y değerleri elde edilir.



Şekil 3.1 Diskriminant analizi sonucunda elde edilen grup dağılımlarının ve ortalama değerlerinin hata bölgeleri ile beraber şematik gösterimi.

Şekil 3.1'de, her iki grubun $y(1)$ ve $y(2)$ ile gösterilen ortalama değerleri ve y_c ile gösterilen grupları birbirinden ayıran bir eşik değeri söz konusudur. Şekilde hata bölgeleri taralı olarak gösterilmiştir.

Ortalama değerler, $y(1)$ ve $y(2)$, elde edilen ağırlıklandırma katsayılarının grupların ortalama özellik değerleri ile çarpılması ile hesaplanır. Eğer bir grubun

diğerine oranla daha büyük olduđu konusunda bir ön bilgi/sezgi yok ise eşik değeri y_c grup ortalamalarının ortalamasıdır.

$$y_c = \frac{y(1) + y(2)}{2} \quad (3.5)$$

Diskriminant modelinin fayda analizi doğru sınıflandırma yüzdesi (doğru gruba atanmış örnek miktarı) ile ölçülebilir. Ancak her veri kümesinin belirli bir kısmı diskriminant modelinden bağımsız olarak şans eseri doğru sınıflandırılabilir. Bu sebeple şans unsuru göz önünde tutularak yani göreceli olarak sınıflandırma başarısı ölçülmelidir. Bu amaçla geliştirilmiş birçok yöntem içinden bu çalışma için maksimum şans kriteri ve orantılı şans kriteri tercih edilmiştir.

Maksimum şans kriteri (Morrison, 1969), $C_{\max}$, seçilen grup için beklenen en düşük doğru sınıflama miktarıdır. Basitçe tüm örneklerin büyük olan gruba atanması ile hesaplanır. Örneğin gruplardaki örnek miktarları 24 ve 76 kabul edersek maksimum şans kriterinin değeri %76 olur. Bu durumda diskriminant modelinin uygulanabilmesi için %76'dan daha yüksek bir sınıflama oranına (doğru grubuna atanan kelimeler/toplam kelime miktarı) sahip olması gerekmektedir.

Gruplar içindeki örnek miktarının aynı olmadığı durumlarda orantılı şans kriteri (C_{pro}) kullanılabilir (Morrison, 1969). Diskriminant modeli grup büyüklüğüne bağlı olarak yapılan sınıflandırmadan daha yüksek bir sınıflandırma başarısı göstermelidir. İki grubun bulunduğu sınıflandırma problemlerinde C_{pro} şu şekilde verilir

$$C_{pro} = p^2 + (1-p)^2 \quad (3.6)$$

Denklemden p birinci gruptaki örnek miktarının tüm örnek büyüklüğüne oranı, $1-p$ ise ikinci gruptaki örnek miktarının tüm örnek kümesine oranını ifade etmektedir. Örneğin, grup büyüklükleri 24 ve 76 iken, orantılı şans kriteri $\%63 (0.24^2 + 0.76^2 = 0.63)$ doğru sınıflandırma olmaktadır.

3.4. İşlev Kelime Tespitinde Diskriminant Analizinin Uygulanması

Diskriminant analizi kelimeleri iki gruba ayırmak için çeşitli bağımsız değişkenlerin/ özelliklerin kullanılmasını temel alan bir yöntemdir. Tez kapsamında bir metinden veya derleminden kolaylıkla elde edilebilecek ve ön işlem miktarı az olan beş özellik kullanılmıştır. Bunlar, kelimenin cümle içindeki sıra numarası, kelime uzunluğu, kelime gözlenme sıklığı (frekansı), belge sıklığı ve eşdizimsel sıklık şeklinde listelenebilir. İzleyen alt bölümde bu özellikler detaylı şekilde tanıtılmaktadır.

Diskriminant analizinin işlev kelime tespitinde uygulanabilirliğini sorgulamak amacıyla yöntemin geçerliliğinin sınanması gereklidir. Çalışmada yöntemin geçerlilik sınaması için k-katlı çapraz geçerlilik testi tercih edilmiştir.

Tez kapsamında işlev kelimelerin tespiti için önerilen yöntem derlem tabanlı bir yöntemdir. Bu yöntemde öncelikle bir talim derlemi üzerinden gerekli ağırlıklandırma katsayıları ve eşik değeri elde edilmelidir. Ardından bir sınama (deney) derleminde bu değerler tatbik edilerek başarımlar ölçülmelidir. İzleyen alt bölümlerde sınıflandırmada kullanılan bağımsız değişkenler, k-katlı çapraz geçerlilik sınaması, diskriminant analizinin uygulanma safhaları olan talim ve sınama safhaları ve değerlendirme yöntemi verilmektedir.

3.4.1. İşlev kelime tespitinde kullanılacak özellikler

Kelime Sıra Numarası: Türkçede cümlelerin temel ögesi yüklemidir. Yüklem bir iş, kılış, oluş, hareket, durum bildiren, haber veren ve cümleyi yargıya bağlayan ögedir. Ögeleri dilin dil bilgisi kuralları dâhilinde genel kurala uygun sıralanmış cümlelerde, yüklem genelde cümle sonunda bulunmaktadır. Diğer ögeler ise önemine göre, vurgulu olan öge yükleme yakın olacak şekilde dizilmektedir². Türkçe’de yüklemi olmayan bir cümlelerin kurulamaması, yüklem içerik kelime olduğu ve cümleye anlam kattığı sonucunu doğurur. Dolayısıyla cümle sonunda bulunan kelime yüksek ihtimalle yüklem ve sonuçta içerik kelime olacaktır.

² <http://tr.wikipedia.org/>

Türkçe cümlelerin büyük bir çoğunluğunun genel dizilim kuralına uygun olarak kurulduğu kabulüyle bir kelimenin içinde bulunduğu cümlede kaçınıcı sırada olduğu yani konum bilgisinin ayırt edici bir özellik olduğu ön görülmüştür. Sıra numarası yükseldikçe bir kelimenin içerik olma ihtimalinin de artacağı kabul edilmektedir. Ancak bir kelimenin içinde bulunduğu cümlede kaçınıcı sırada olduğunun tespiti, cümle uzunluğundan bağımsız olarak yapılmalıdır. Aksi takdirde uzun cümlelerde cümle ortasında bulunan bir işlev kelime, kısa bir cümlelerin yüklemi ile aynı sıra numarasına sahip olacak ve özelliğin ayırt ediciliği azalacaktır. Bu sebeple sıra numarası belirlenirken kelimenin cümle içindeki pozisyon değeri cümle uzunluğuna bölünmektedir. Örneğin 5 kelimelik bir cümlede 3. kelime için sıra numarası $3/5$ olarak hesaplanarak, bu değer cümle uzunluğundan bağımız hale getirilmiştir.

Kelime Uzunluğu: Halihazırda kullanılan işlev kelime listeleri incelendiğinde pek çok işlev kelimenin (“*ve, bir, bu, gibi, ile, şey, de, da, ki*”) az sayıda harften oluştuğu görülmektedir. Çok sıklıkla kullanılan bir kelimenin diğer kelimelere oranla daha kısa olması en az gayret prensibi dikkate alındığında beklenen bir durumdur (Zipf, 1949). Ayrıca işlev kelimeler anlama katkıda bulunmayacakları için yapım ekleri olarak yeni kelimeler üretmeyeceklerdir. İçerik kelimeler ise kökleri az sayıda harften oluşsa dahi yapım ekleri olarak yeni kelimeler türetme, yeni anlamlar taşıma dolayısıyla da harf sayılarını arttırma eğilimindedirler. Bu sebeplerle kelimelerin gövde ve köklerinin eklerinden ayrıştırılmadığı bir derlemde, işlev kelimelerin içerik kelimelere oranla daha az harf sayısına sahip oldukları kabul edilerek, kısa kelimelerin daha yüksek olasılıkla işlev kelime olacağı öngörülmüş ve kelime uzunluğu özelliği analize dâhil edilmiştir.

Kelime Gözlenme Sıklığı: İşlev kelimeler anlam taşımaktan çok cümle yapısının kurulması için kullanıldıkları için metinlerde çok sıklıkla gözlenmektedirler. Bu sebeple farklı dillerde yapılan çalışmalar sonucunda gözlenme sıklığı yüksek olan kelimelerin işlev kelime olduğu araştırmacılar tarafından kabul edilmektedir.

Çalışma içinde elde edilen gözlenme sıklığı değerlerinin derlem büyüklüğünden etkilenmemesi için belli bir pencere dâhilinde sıklık bilgisi ölçülmüştür. Pencere olarak talim derlemi için 1000 kelime, deney derlemi için 500 kelime alınmıştır.

Böylece talim derlemindeki her bir kelime için kendinden önceki 500 ve kendinden sonraki 500 kelime içinde kaç kere gözleendiği belirlenmiş ve sıklık değeri olarak kabul edilmiştir.

Belge Sıklığı: Farklı konular içeren belgelerde gözlenen kelimeler tf-idf gibi çeşitli yöntemlerde işlev kelime olarak kabul edilmektedir. Tez içerisinde bu özellik işlev kelimeleri içerik kelimelerden ayıran bir özellik olarak kabul edilerek bir kelimenin en az bir kere gözleendiği belge miktarının derlem içindeki toplam belge miktarına oranı belge sıklığı olarak analize dâhil edilmiştir.

Eşdizimsel Gözlenme Sıklığı: Eşdizim (Ing. collocation) rastgele olamayacak sıklıkta yan yana gözlenen iki veya daha çok kelimedenden oluşan kelime gruplarıdır (Van der Vouden, 1997). Oxford Eşdizim Sözlüğündeki tanımıyla bu kelimeler doğal, akıcı yazım ve konuşma üretmek amacıyla dil içinde bir araya gelmektedirler. Türkçe eşdizimlere örnek olarak “peşin hükümlü, kesin karar, idare etmek, kaba saba, geç kalmak, evrak kalemi” gibi kelime ikilileri verilebilir (Sarıkay, 2006).

İki kelimedenden oluşan eşdizimlerin tespitinde kullanılan en basit yöntem derlem içinde sıklıkla yan yana gözlenen kelimelerin tespitidir. Ancak yöntem test edildiğinde çok sıklıkla gözlenen ikililer içinde işlev kelimelerin de olduğu gözlenmiştir (Manning and Schütze, 1999). İşlev kelimelerin dilin içerisinde çok sıklıkla kullanılıyor olması onların herhangi bir kelime ile birlikte gözlenme olasılıklarını arttırmaktadır. Aynı zamanda farklı kelimelerle bir arada gözlenme olasılıklarını da yükseltmektedir.

Bu fikirlerden yola çıkarak sıklıkla rastlanan ikililer, gerçek eşdizimler ve işlev kelimeler içeren hatalı eşdizim adaylarından oluşur denilebilir. Önerilen yöntem içinde bu bilgiden faydalanılmaktadır. Her kelimenin kaç farklı eşdizim adayı yani sıklıkla gözlenen ikili içinde yer aldığı bilgisi bir özellik olarak analize dâhil edilmiştir. Çok sayıda aday içinde rastlanan kelimeler işlev, az sayıda aday içinde yer alan kelimeler ise içerik olarak kabul edilmektedir. Örneğin, Çizelge 3.1’de OSTAD derleminden elde edilen gözlenme sıklığı en yüksek olan ilk 20 kelime ikilisi listelenmiştir (OSTAD derleminin özellikleri Bölüm 3.4.3’de

verilmiştir). Bu liste değerlendirildiğinde “bir” işlev kelimesinin listede yer alan 7 farklı ikilide, “büyük”, “zaman” ve “sordu” gibi içerik kelimelerin ise bir tek ikili de yer aldığı gözlenmektedir. Benzer şekilde diğer işlev kelimeler farklı ikililerde sıklıkla rastlanmaktadır.

Bu özelliğin derlem büyüklüğünden bağımsız hale getirilmesi için bir kelimenin rastlandığı farklı eşdizim aday sayısı toplam eşdizim aday sayısına bölünmüştür. Örneğin toplam 400 adet aday eşdizimin olduğu bir derlemde “ve” kelimesi 30 farklı aday içinde gözlensin. Bu durumda kelimeye ait farklı eşdizim adayı içinde gözlenme miktarı 30/400 olarak hesaplanır.

Çizelge 3.1. OSTAD derleminde sıklıkla yer alan kelime ikilileri

Sıra Numarası	Gözlenme Sıklığı	Kelime İkili	Sıra Numarası	Gözlenme Sıklığı	Kelime İkili
1	49	ya da	11	20	sonra da
2	41	bir şey	12	19	diye sordu
3	31	ben de	13	19	belki de
4	27	o da	14	16	her şey
5	24	bir an	15	16	değil mi
6	23	ne kadar	16	16	daha çok
7	23	böyle bir	17	15	o zaman
8	23	bir süre	18	15	büyük bir
9	22	bir de	19	15	başka bir
10	21	o kadar	20	14	çok iyi

Çalışmada derlemlerde sık rastlanan ikililerin belirlenmesi için eşik değeri 4 olarak kabul edilmiştir. Yani derlemlerde en az 4 kere rastlanan ikililer eşdizim adayı olmaktadır.

3.4.2. Geçerlilik testi

Diskriminant analizinin kelimeleri iki gruba ayırmadaki geçerliliğinin tespiti için K-katlı çapraz geçerlilik testi (İng. K-fold cross-validation) uygulanmıştır. Bu testte, veri seti K adet alt kümeye bölünüp, K-1 adet alt veri kümesi talim derlemi olarak, 1 adet veri kümesi ise deney derlemi olarak kullanılır. Test içinde tüm alt kümeler bir kez deney derlemi olarak alınıp, K kez işlem tekrarlanarak sınıflamadaki başarımlar değerlendirilir. K adet başarımların ortalaması testin geçerlilik değerini verir.

3.4.3. Talim ve Sınama

İşlev kelimelerin tespitiinde önerilen diskriminant analizi yönteminin uygulanabilmesi için öncelikle diskriminant fonksiyonunda kullanılacak katsayıların bir derlem (talim derlemi) üzerinden elde edilmesi gereklidir. Çalışmamızda işaretli bir derlem olan OSTAD (ODTÜ-Şabancı TALim Derlemi) (Ofłazer et al., 2003; Atalay et al., 2003) talim derlemi olarak seçilmiştir.

İşlev kelime tespitinde talim aşamasında kullanılan OSTAD derleminin cümle sonu işaretlemesi elle gerçekleştirilmiştir. Analiz başarımının, cümle sonu işaretlemesi başarımından etkilenmemesi için otomatik yöntemler tercih edilmemiş olsa da ileriki çalışmalarda kullanılmaları mümkündür. Literatürde cümle sonu işaretlemesi konusunda hem kural tabanlı hem de makine öğrenmesi yaklaşımını taşıyan pek çok çalışma bulunmaktadır (Dinçer and Karaoğlan, 2005; Tür et al., 2003; Reynar and Ratnaparkhi, 1997; Riley, 1989; Palmer and Hearst, 1997; Mikheev, 2000).

İşlev-içerik kelime ayrıştırılması çalışmasında talim derlemi olarak OSTAD'ın tercih edilmesinin temel nedenleri, kelime türleri ve cümle sonlarının elle işaretlenmiş olması; farklı konularda metinleri içermesi sebebiyle dilin çeşitliliğini simgelemekte daha başarılı olacağı öngörüsüdür.

OSTAD derleminin incelenmesi aşamasında bir takım metinlerin tekrarlarına rastlanmıştır. Tekrar metinlerin ayıklanması sonucunda OSTAD (toplam 46532 kelime, 33 belge) derleminden çizelge 3.2'de verilen istatistikler elde edilmiştir. OSTAD derleminden çekilen bilgilerde herhangi bir değişiklik yapılmadan yani olası tür bilgisi ve işaretleme hataları ayıklanmadan çalışma içinde kullanılmıştır.

İşlev kelime tespiti işleminin talim aşaması talim derleminde işlev kelimelerin etiketlenmesini gerektirir. Tez içerisinde işlev kelimelerin etiketlenmesi aşamasında kelime türü bilgisi kullanılmıştır. İsim, fiil ve sıfat olarak işaretlenmiş olan kelimeler anlam taşımaları yani içeriğe, cümlelerin

aktardığı enformasyona büyük oranda katkılarının olmaları sebebiyle içerik olarak etiketlenmiştir. Diğer bütün türler ise işlev kelime olarak kabul edilmiştir.

Çizelge 3.2. OSTAD derleminin genel özellikleri

Kelime Türü	Rastlanma miktarı %	Örnekler
İsim	43.26%	ömür, kahvenin, muammer, düğün
Fiil	24.36%	söyledi, dedi, bulundum
Sıfat	8.27%	iri, genç, yavaş, dar
Bağlaç	4.92%	ama, da, ki, ile, ve, ancak
Zamir	4.65%	hepimiz, kendileri, onu, onun, ben
Zarf	4.64%	biraz, sanki, şimdi, nasıl, çok, daha
Belirteç	4.35%	bir, o, hiçbir, her, bu
Edat	2.72%	diye, gibi, sonra, kadar, doğrusu
Sayı	2.05%	iki, ikimiz, beşinci, yirmilerindeki
Soru eki	0.49%	mı, musunuz, mi
Seslenme Edatı	0.23%	eee, hadi, ah, hayır, yok, işte
Tekrar	0.04%	pırıl, gürül, tefek
Belirsiz	0.02%	katana, sezen'in, koha

Talim aşamasında elde edilen ağırlıklandırma katsayıları ve eşik değeri sınama derlemi üzerinde tatbik edilmiş ve başarımlar değeri ölçülmüştür. Talim safhasında elde edilen ağırlıklandırma katsayıları ve eşik değeri Bölüm 3.5 Sonuçlarda (Çizelge 3.5) verilmiştir. Bu deney aşaması deney derleminin tümü veya derlem içinde yer alan her belge için ayrı bir aday işlev kelime listesi üretebilir.

Çalışmada sınama derlemi olarak farklı konular içeren 8 adet belge/metin kullanılmıştır. Bu belgelerin cümle sonu işaretlemesi elle yapılmıştır. Çizelge 3.3 bu belgeler hakkında bir takım istatistikleri vermektedir. Yöntemin sınanması için her belge için üretilen listeler diğerlerinden bağımsız olarak değerlendirilmiştir. Benzer şekilde bu sınama metinleri kelime sıklık ve tf-idf yöntemlerine tabi tutularak bu iki yöntem için de sonuçlar elde edilmiştir.

Çizelge 3.3. İşlev-içerik kelime ayrımı yapılan örnek Türkçe metinlerin genel özellikleri

Metin No	Belge Konusu	Cümle Sayısı	Kelime Sayısı	Kelime Dağılımı
1	Sigara Bağımlılığı	46	510	363
2	Hayvanlar	82	928	592
3	Hemofili Hastalığı	105	1407	818
4	Türkiye'de Bilim	108	1497	955
5	Uyuşturucu Madde Bağımlılığı	161	1734	1050
6	Elektronik Kitap	118	1762	964
7	Alman Devrimi	95	2639	1385
8	Minyatür Sanatı	174	2901	1570

3.4.4. Değerlendirme

Tez kapsamında geliştirilen yöntem ve bu yöntemin kıyaslanacağı kelime sıklık ve tf-idf yöntemlerinin değerlendirilmesi için duyarlılık ölçeği kullanılmıştır. İşlev kelime tespiti çalışmalarında, duyarlılık, yöntem tarafından belirlenen doğru işlev kelime miktarının aday işlev kelime listesinin büyüklüğüne oranı olarak tanımlanabilir. Elde edilen aday işlev kelime listelerinde doğru işlev kelimelerin işaretlenmesi için derlemde yer alan ve yöntemlerin listelerinde yer alan işlev kelimeler (örneğin “bir”, “bu”, “şu”, “kendi” gibi) ve bunların farklı yüzey formları, sayılar (örneğin “iki”, “1860” gibi) ve harflerden oluşan bir baz işlev kelime listesi kullanılmıştır (Bkz. Ek 1).

Duyarlılık değerlerinin sunulması esnasında her üç yöntem için tek bir noktada (örneğin ilk 100 aday kelime için) duyarlılık değeri vermek yerine tüm noktalardaki sonuçları gösteren duyarlılık eğrileri çıkartılmıştır. Kelime sıklık ve tf-idf yöntemleri sıralı aday listeleri üreten yöntemlerdir. Bu yöntemlerin ürettikleri listelerin belirli bir kısmı için değil her anı için duyarlılık ölçülmüştür. Yani duyarlılık sadece ilk n aday için ölçülmemiştir, n değerinin 1 ile N arası tüm değerleri için duyarlılık hesaplanmıştır. Bu N değeri aday listeler için bir eşik değeridir. Kelime sıklık ve tf-idf yöntemleri için eşik değer belirlemek daha önce de bahsedildiği üzere oldukça tartışmalı bir konudur. Bu sebeple bu çalışmada N değeri diskriminant analizinin ürettiği aday işlev kelime sayısı olarak kabul edilmektedir. Böylelikle her yöntem ilk (en iyi) N adayı dikkate

alınarak kıyaslanmıştır. Diğer yöntemlerden farklı olarak diskriminant analizi sıralı bir liste üretmemektedir. Yani ürettiği listedeki herhangi bir aday diğerinden daha öncelikli olarak işlevdir denilemez. Diğer yöntemler ile kıyaslanabilirliği sağlamak amacıyla diskriminant analizinin ürettiği listedeki adaylar belge sıklık değeri azalacak şekilde sıralanarak duyarlılık eğrileri hazırlanmıştır.

3.5. Sonuçlar

K-katlı çapraz geçerlilik testi için talim derlemi yaklaşık aynı büyüklükte 10 gruba ($k=10$) bölünmüştür. Çizelge 3.4 OSTAD derleminden oluşturulan bu 10 grup ile ilgili bir takım istatistikleri vermektedir.

10-katlı çapraz geçerlilik testi sonucunda ortalama %84.89 başarımlı oranı (doğru sınıflandırma yüzdesi) elde edilmiştir (standart sapma değeri=0,039). Derlemde işlev kelimeleri ayırt etme konusunda etkin özellikler sırasıyla kelime gözlenme sıklığı, eşdizimsel sıklık, belge sıklığı, sıra numarası ve kelime uzunluğu olarak gözlenmiştir. Bağımsız değişkenlerin teker teker analize dâhil edilmesi ile gerçekleştirilen aşamalı test sonucunda, tüm bağımsız değişkenlerin analiz içinde yer alması gerektiği görülmüştür.

Çizelge 3.4. K- katlı çapraz geçerlilik testinden kullanılan gruplara ait istatistikler.

Alt grup no	Cümle Sayısı	Kelime Sayısı
1	743	4649
2	497	4653
3	386	4652
4	453	4657
5	740	4647
6	844	4654
7	678	4661
8	587	4656
9	392	4648
10	345	4655

Yöntemin geçerliliğinin sınamasından sonra ilgili bağımsız değişkenler analize dâhil edilerek OSTAD derlemi kullanılarak ağırlıklandırma katsayıları ve eşik değeri belirlenmiştir. Çizelge 3.5’de ağırlıklandırma katsayıları, eşik değeri, doğru sınıflandırma oranı, maksimum ve orantılı şans kriterleri verilmiştir. Eşik

değer (y_c) iki grubun yaklaşık aynı miktarda kelime içerdiği kabulüyle bu gruplar için belirlenen eşik değerlerin ortalaması olarak hesaplanmıştır. Derleminden elde edilen doğru sınıflama değeri maksimum ve orantılı şans kriteri ile kıyaslandığında daha büyük olduğu yani yöntemin şans eseri yapılacak bir sınıflamadan daha başarılı olduğu görülmektedir.

Çizelge 3.5. Talim derleminden elde edilen diskriminant analizi sonuçları

Ağırlıklandırma Katsayısı	
Kelime Sıklığı	57,04
Eşdizimsel Sıklık	-4,29
Kelime Uzunluğu	-0,10
Kelime Sırası	-0,39
Belge Sıklığı	3,31
Sabit	-0,04
Eşik Değer y_c	0,56
Doğru Sınıflama Oranı	86.30%
C_{max}	75.89%
C_{pro}	63.41%

Sınama metinleri (derlemi) talim derleminden elde edilen katsayılar ve eşik değer ile diskriminant analizine tabi tutulduğunda elde edilen aday işlev kelime sayıları belge bazında Çizelge 3.6'da verilmektedir. Ek 2'de diskriminant analizinin ürettiği listede yer alan ilk 50 aday işlev kelime listesi bulunmaktadır.

Çizelge 3.6. Sınama derleminde diskriminant analizi sonucunda elde edilen aday işlev kelime miktarları

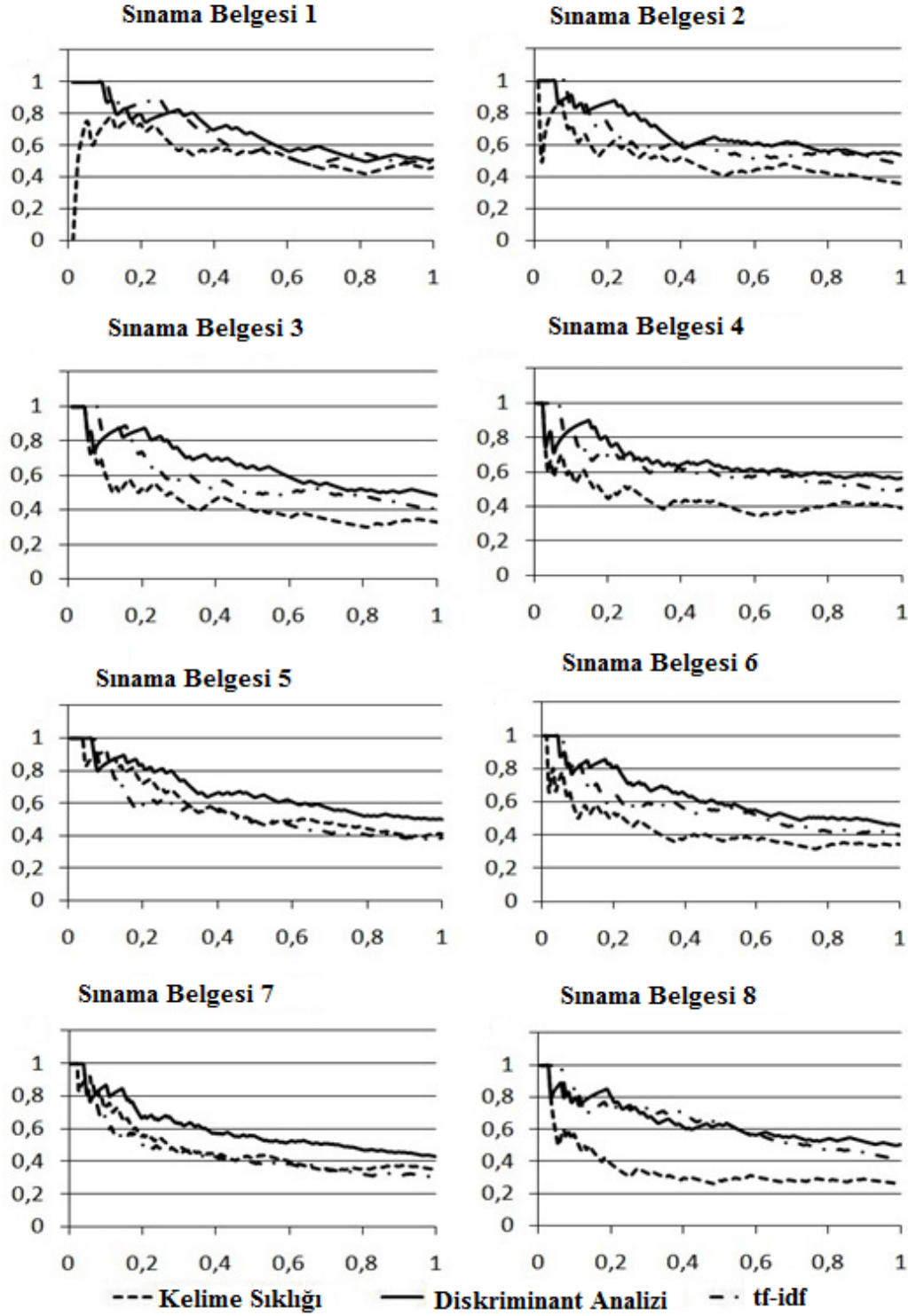
Belge No	Belge Konusu	Kelime Sayısı	Kelime	Aday işlev
1	Sigara Bağımlılığı	510	363	76
2	Hayvanlar	928	592	110
3	Hemofili Hastalığı	1407	818	116
4	Türkiye'de Bilim	1497	955	134
5	Uyuşturucu Madde Bağımlılığı	1734	1050	126
6	Elektronik Kitap	1762	964	158
7	Alman Devrimi	2639	1385	227
8	Minyatür Sanatı	2901	1570	142

Çizelge 3.6'dan da görüldüğü üzere aday işlev kelime sayısı metin uzunluğundan bağımsız değerler olmuştur. Çizelge 3.7'de tüm yöntemlerde diskriminant analizinin ürettiği işlev kelime miktarı kadar kelime için ölçülen duyarlılık değerleri gösterilmiştir.

Şekil 3.2’de kelime sıklığı, tf-idf ve diskriminant analizi yöntemlerinin her bir sınama metni için duyarlılık grafikleri verilmiştir (Ek 3 ve Ek 4’de kelime sıklık ve tf-idf yöntemlerinin ürettiği ilk 50 adaya ait listeler bulunmaktadır). Grafiklerde yatay eksen tamamlanan aday işlev kelime miktarının toplam aday işlev kelime miktarına (N) oranını simgelemektedir. Grafiklerden de görüldüğü üzere tüm sınama belgelerinde/metinlerinde gözlenme sıklığı yöntemi en düşük duyarlılık değerlerini üretmektedir. Diskriminant analizi ise toplam aday miktarının ilk çeyreğinde tf-idf yöntemi ile kabaca aynı değerleri üretse dahi aday listesinin kalan kısmı için daha yüksek duyarlılık değerleri göstermektedir. Ayrıca tf-idf ve kelime sıklık yöntemleri için aday listelerinin grafiklerde duyarlılık değeri gözlenen kısmını takip eden kısmında duyarlılık değerinin daha hızlı bir şekilde düştüğü gözlenmiştir.

Çizelge 3.7. Sınama belgelerinde diskriminant analizi sonucunda elde edilen duyarlılık değerleri (İşlev kelime miktarı diskriminant analizinin ürettiği işlev kelime miktarı olarak kabul edilmiştir)

Belge No	Belge Konusu	İşlev Kelime Sayısı	Duyarlılık		Diskriminant Analizi
			Gözlenme Sıklığı	tf-idf	
1	Sigara Bağımlılığı	76	46,05%	48,68%	51,32%
2	Hayvanlar	110	35,45%	49,09%	53,64%
3	Hemofili Hastalığı	116	32,76%	41,38%	48,28%
4	Türkiye’de Bilim	134	38,81%	50,00%	56,72%
5	Uyuşturucu Madde Bağımlılığı	126	41,27%	38,89%	50,00%
6	Elektronik Kitap	158	34,18%	39,87%	45,57%
7	Alman Devrimi	227	35,68%	30,40%	43,17%
8	Minyatür Sanatı	142	26,06%	40,85%	50,70%



Şekil 3.2. Diskriminant analizi, gözlenme sıklığı ve tf-idf yöntemleri için deney belgelerinden elde edilen duyarlılık grafikleri (yatay eksen tamamlanan aday işlev kelime miktarının toplam aday işlev kelime miktarına olan oranı, dikey eksen ise duyarlılık değerini ifade etmektedir).

4. EŞDİZİMLERİN TESPİTİ

4.1. Eşdizimlerin Tespiti Problemi

Eşdizim, rastgele olamayacak sıklıkla bir arada gözlenen geleneksel kelime birlikleridir. Eşdizim kavramı ilk kez İngiliz dilbilimci J.R. Firth tarafından 1951 yılında yayınlanan “Modes of Meaning” adlı kitapta yer almıştır. Firth bir kelimenin ancak, kendisine eşlik eden kelime ile değerlendirilebileceğini vurgulamış ve “Verilen bir kelimenin eşdizimleri bu kelimenin alışlagelmiş veya geleneksel pozisyonlarıdır” şeklinde bir tanım ortaya koymuştur. Daha sonra Firth’in öğrencisi olan Sinclair eşdizimi, bir metin içerisinde iki veya daha çok kelimenin kısa mesafede yer alması şeklinde tanımlamıştır (1991). Hoey ise istatistiksel bir yaklaşımla eşdizimlere “Kendi bağlamında rastlantısal olamayacak miktarda bir arada görülen sözcükler arası ilişkiye verilen isim” demiştir (1991). Literatürde eşdizim tanımının farklı şekillerde yapıldığı birçok çalışma yer almaktadır (Haas, 1966; Halliday, 1961; Lyons, 1966; McIntosh, 1966; Van Buren, 1967; Leech, 1974; Newmark, 1978; Benson, 1990; Smadja, 1993).

Eşdizimler keyfidirler, diğer bir deyişle eşdizimlerin hangi sebeple ve ne şekilde oluştuğunu tanımlayan kesin kurallar yoktur ve eşdizimler çoğunlukla anlamsal olarak açık değildirler. Ancak yine de yazım ve konuşmada önemli bir anlamsal etkiye sahiptirler. Bu sebeple eşdizimlerin belirlenmesi pek çok doğal dil işleme çalışmasında; doğal dil üretme, makine çevirisi, anlam belirsizliğinin giderilmesi, kelime türü tespiti, bilgi çıkarsama ve derlem dilbilim (Bisht et al., 2006; Manning and Schütze, 1999); önemli bir konudur. Bu denli geniş bir uygulama alanına hizmet eden eşdizimlerin belirlenmesi amacıyla birçok istatistiksel ve kural tabanlı yöntemler geliştirilmiştir. Kural tabanlı yöntemler kelime türü işaretlenmiş veriler üzerinde çalışan ve metin üzerinde bir grup öncül işlemin gerçekleşmesini gerektiren yöntemlerdir. İstatistiksel yöntemler ise bir çeşit gözlenme sıklığı (İng. frequency of occurrence) bilgisini temel alarak verilen bir derlemde eşdizimleri tespit eden yöntemlerdir. En sıklıkla kullanılan istatistiksel teknikler gözlenme sıklığı, karşılıklı enformasyon (mutual information) (Church and Hanks, 1990) ve log-olabilirlik ki-kare χ^2 , t-testi gibi hipotez testleridir. Smadja’nın Xtract’ı (1993), Kita vd. yöntemi (1994),

Shimohata vd. teknikleri (1997) de literatürde araştırmacıların adları ile anılan önemli çalışmalardandır. Çeşitli eşdizim yöntemleri ile ilgili olarak daha detaylı bilgiler Manning ve Schütze çalışmasında bulunabilir (1999).

Tez kapsamında öncelikle Türkçe bir derlem üzerinde eşdizimlerin otomatik tespiti için literatürde yer alan bir takım istatistiksel teknikler uygulanmıştır. İlk olarak tekniklerin ürettikleri sonuçlar kendi içinde değerlendirilerek Türkçe'nin yapısına ve tez amacına uygun olan yöntemler belirlenmiştir. Daha sonra bu tekniklerin başarımları bilgi geri getirim alanında sıklıkla kullanılan duyarlılık (İng. precision) ve anma (İng. recall) ölçütleri ile değerlendirilmiş ve birbiri ile kıyaslanmıştır. Eşdizim tespiti için derlemin hem yüzeysel hem de gövdelenmiş hali üzerinde çalışma yürütülerek gövdelemenin eşdizim tespitindeki etkisi araştırılmıştır. Her ne kadar farklı dillerde, İngilizce, Rusça, Çince, Fransızca, vd. gibi, eşdizimlerin saptanması üzerine hâlihazırdaki yöntemlerin kıyaslanmasına dair çalışmalar yapılmış olsa da Türkçe üzerine bu kapsamda yapılmış bir araştırma bilimiz dâhilinde yoktur.

Tez içerisinde eşdizimlerin tespiti için kelimeler arası bağların ve kelime anlam bütünlüğünün değerlendirilmesine dayanan “eşdizim eğilimi yöntemi” geliştirilmiştir. Yöntem, Türkçe derlem üzerinde uygulanarak etkisi incelenmiştir. Geliştirilen yöntem hâlihazırda literatürde yer alan yöntemlerin ürettiği eşdizim listeleri üzerine tatbik edilmiş ve ilgili sonuçları iyileştirdiği gösterilmiştir.

İzleyen bölümlerde sırasıyla eşdizim tanımı verilmekte, Türkçe üzerine gerçekleştirilen eşdizimlerin tespiti çalışmaları değerlendirilmekte, literatürde yer alan eşdizim tespit yöntemleri tanıtılmakta, tez kapsamında önerilen eşdizim eğilimi yöntemi sunulmakta, deneysel çalışma için kullanılan derlemeler ile ilgili bilgiler verilmekte, deneysel hazırlık çalışmaları ve deney sonuçları verilmekte ve tartışma bölümü yer almaktadır.

4.2. Eşdizim Özellikleri

Literatürde yer alan pek çok farklı tanımdan da anlaşılacağı üzere aslında eşdizim oluşturmak için kesin kurallar yoktur ve dil kendi gelişimi içinde eşdizimleri oluşturur.

Her ne kadar araştırmacıların ortak olarak kabul ettikleri bir eşdizim tanımı olmasa da eşdizim özellikleri için farklı çalışmalarda ortak olarak rastlanan özellikler şöyle tanımlanabilir:

Eşdizimler sıklıkla yinelenir: Yinelenme (İng. recurrence) özelliği eşdizimlerin diğer sözcük birliklerinden ayırt edilmelerini sağlayan ve ölçümü en kolay olan özellikleridir. Bu sebeple, eşdizim tespit tekniklerinin neredeyse tümünde eşdizimin rastlanma sıklığı ile bir bilgi yer almaktadır (Bisht et al., 2006; Church and Hanks, 1990; Smadja, 1993; Hindle, 1990; Dunning, 1993).

Eşdizimler keyfi ve dil bağımlıdır: Eşdizimler herhangi bir kurala bağlı olmaksızın gelişmişlerdir. Dil içinde hangi kelimelerin eşdizim oluşturacağı hangilerinin oluşturmayacağı; bir kelimenin milyonlarca kelime içinden hangisini seçip bir eşdizim oluşturacağı konusunda bilinen bir kural yoktur. Örneğin “kör talih” Türkçe’de sıklıkla kullandığımız eşdizimlerdenidir. Ancak bu eşdizimde “kör” kelimesi yerine eşanlamlı olmasına rağmen niçin “âmâ” kelimesini 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 kelime veya kelime grubu olarak tanımlanır. Bu kelime veya kelime grubu, cümle veya cümle ögesi olarak görev yapar. Özellikle anlam belirsizliğinin giderilmesi, kelime öğelerinin saptanması makine çevirisi gibi anlam bütünlüğünün önemsendiği çalışmalarda birim bloklar önemlidir. Örneğin İngilizcedeki “lady killer” eşdizimi bir bütün, bir blok olarak kabul edilmeyerek Türkçeye “kadın katili” şeklinde

çevrilebilir. Oysa bu eşdizim bir birim olarak kabul edildiğinde “çapkın” kelimesine denk gelir.

Doğal dil içerisinde birim bloklar oluşturan kelime grupları içinde eşdizimlerin yanı sıra deyimler ve atasözleri de yer alır. Özellikle deyim ve eşdizim kelimeleri pek çok çalışmada eş anlamlı olarak kullanılsa da, ki tez kapsamında da benzer bir durum söz konusudur, aslında tam olarak aynı kavramı ifade etmezler. Deyimleri oluşturan kelimeler biraraya geldiklerinde kelime anlamlarının tamamıyla dışına çıkarak yeni bir anlam üretirler. Örneğin “etekleri zil çalmak” deyiminde fiilen eteklerin zil çalması söz konusu değildir. Eşdizimlerde ise kelimeler bir araya gelirken kendi anlamlarını tam olarak yitirmezler. Örneğin “kör talih” eşdiziminde talih kelimesi anlamını yitirmezken kör kelimesi anlam değişikliğine uğramıştır. Bu durumun ifadesi olarak, Manning ve Schütze (1999) çalışmalarında eşdizimleri sınırlı bütünlüğe (limited compositionality) sahip kelime birlikleri olarak tanımlamaktadır. Tez kapsamında ise eşdizim kavramı deyimleri de kapsayan bir üst küme olarak kabul edilmektedir.

Eşdizimler alan bağımlıdır: Eşdizimler dil bağımlı oldukları gibi aynı zamanda spor, sanat, kültür, bilim vb. gibi alanlar içinde de özelleşirler. Smadja çalışmasında denizcilik alanını örnek göstermiştir (1993). “Islak 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, tez kapsamında eşdizim olarak kabul edilen kelime 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.)

Eşdizim oluşturan kelimelerin iki veya daha fazla olabileceği ve kendi aralarına farklı miktarda kelime alabileceği kabul edilmiştir (Smadja,1993; Shimohata et al., 1997). Türkçe metinler üzerine yaptığımız inceleme sonucunda kendi arasına kelime alan veya ikiden fazla kelime ile oluşturulan eşdizimlerin deney dışı bırakılmasına, bu tip eşdizimlerin yakalanması için sarf edilecek emek miktarının sonuçtaki başarımla eş değer olmayacağına karar verilmiştir. Bu sebeple çalışma, derlemde yan yana gözlenen iki kelimedenden oluşan eşdizimlerin tespiti ile sınırlandırılmıştır. Bundan böyle eşdizim yan yana bulunan iki kelimedenden oluşan ve yukarıda belirtilen özellikleri taşıyan kelime birlikleri olarak anlaşılmalıdır.

4.3. Eşdizim Tespit Yöntemleri

Eşdizimlerin tespitinde kural tabanlı ve istatistiksel yöntemler mevcuttur. Kural tabanlı yöntemler bir sorgu kümesi vasıtasıyla eşdizimlerin tespitini amaçlayan yöntemlerdir. Bu yöntemler sözcük türü gibi bir takım özellikleri etiketlenmiş yani öncül bir işlemde geçmiş derlemlerde çok miktarda sorgunun işletilmesini gerektirirler. Bu sebeple zaman karmaşıklıkları yüksektir. İstatistiksel teknikler ise öncül işlemleri en aza indirgeyerek zaman karmaşıklığını azaltmayı öngören tekniklerdir. İzleyen alt bölümlerde literatürde sıklıkla kullanılan ve/veya dilden bağımsız olarak uygulanabilen istatistiksel teknikler: gözlenme sıklığı, karşılıklı enformasyon, ortalama-varyans, Smadja yöntemi, Shimoto-Sugio-Nagata yöntemi ve hipotez testleri tanıtılmaktadır.

4.3.1. Gözlenme sıklığı

İki veya daha fazla kelimenin bir arada gözlenme sayısının (sıklık) eşdizim tespitinde kullanılması en basit ve en eski yöntemdir. Yöntem uygulanırken derlemde yan yana bulunan kelimelerin veya bir pencere dâhilinde bir arada bulunan kelimelerin sıklık değerleri ölçülür. Derlemdeki bu kelime 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 eşdizimleri, diğer kelime birliklerinden ayıran eşik değerin belirlenmesidir. Yöntemin dezavantajı ise çok sıklıkla

gözlenen eşdizim adayları içinde işlev kelimelere (örneğin bir, bu, ve, şey gibi) rastlanmasıdır (Manning and Schütze, 1999). Bu sebeple sıklık değerine bağlı olarak elde edilen listeler çeşitli filtrelerden geçirilir (örneğin sözcük türü filtresi) (Justeson and Katz, 1995).

4.3.2. Karşılıklı enformasyon

Karşılıklı enformasyon, enformasyon teorisindeki tanımıyla, iki rassal değişkenin karşılıklı bağımlılıklarını gösteren bir değerdir. İki değişkenin bir arada görülme olasılığının ayrı ayrı görülme olasılıklarına bölünmesi ile elde edilen değer iki tabanında logaritması alınarak bu bağımlılığın bit cinsinden değeri ifade edilir.

Hesaplamalı dil bilim çalışmalarında, eşdizim oluşturan kelimelerin bir arada bulunuşlarının tesadüfî olmadığı fikrinden yola çıkarak kelimeler arası bağımlılıkların miktarını saptamak için karşılıklı enformasyon yöntemi kullanılmaktadır (Church and Hanks, 1990; Hindle, 1990). Bu yöntemde bir arada görünen iki kelime için her kelimenin kendi başına görülme olasılığı ile birlikte görülme olasılığı arasındaki ilişki hesaplanır. Bu ilişkiye dayanarak eşdizimlilik kararı verilir. Denklem 4.1’de x ve y kelimeleri için karşılıklı enformasyon değeri yer almaktadır.

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x) \cdot P(y)} \quad (4.1)$$

x ve y kelimeleri birbirinden bağımsız ise bu iki kelimenin bir arada gözükme olasılığı, kelimelerin gözlenme olasılıklarının çarpımına eşittir ($P(x, y) = P(x) \cdot P(y)$). Dolayısıyla eşdizim oluşturmayan x ve y kelimeleri için karşılıklı enformasyon değeri sıfır olmaktadır ($I(x, y) = 0$). Bu iki kelimenin bağımlı olduğu yani eşdizim oluşturduğu durumda ise karşılıklı enformasyon değeri sıfırdan uzaklaşacaktır. Bu sebeple herhangi bir kelime birliği için karşılıklı enformasyon değeri sıfır değerinden ne kadar uzaklaşır ise bu birliklerin eşdizim özelliğinin o denli arttığı kabul edilir.

Yöntemin uygulanmasında gözlenme sıklığı yöntemine benzer şekilde tüm derlemde kelime birlikleri için karşılıklı enformasyon değerleri ölçülür ve bu değerler azalacak şekilde kelime birlikleri listelenir. Listede yüksek karşılıklı enformasyona sahip olan birlikler eşdizim olarak kabul edilir.

4.3.3. Hipotez testleri

Bir kelime birliğinin eşdizim olup olmadığının tespiti için kelimelerin bir arada bulunuşlarının şans eseri olmadığının ispatı gereklidir. Kelimeler arası bağımlılığı göstermek için çoğunlukla kelimeler arası bağımsızlık test edilir. Hipotez testlerinde, sıfır hipotezi olarak kelimelerin birbirinden bağımsız olduğu kabul edilir ve sıfır hipotezi reddedilmeye çalışılır. Eşdizim tespiti için sıklıkla kullanılan hipotez testleri takip eden bölümlerde kısaca anlatılmaktadır.

4.3.3.1. Dunning log-olabilirlik testi (Dunning's log-likelihood test)

Log-olabilirlik yöntemi Dunning tarafından önerilen bir hipotez testidir (Dunning, 1993). Eşdizim tespitinde, w_1w_2 kelime ikilisinin gözlenme sıklığı için iki alternatif tanım bulunmaktadır:

$$\text{Hipotez 1 : } P(w_2 / w_1) = p = P(w_2 / \neg w_1)$$

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

Hipotez 1, w_2 kelimesinin gözlenme durumunun, daha önce gözlenen w_1 kelimesinden bağımsız olduğunu ifade eder. Hipotez 2 ise w_1 kelimesinin gözlendiği ve gözlenmediği durumlarda w_2 kelimesinin gözlenme ihtimalinin farklılaştığını belirtir. Eğer hipotez 1 kabul edilirse, kelime ikilisi bir eşdizim değildir, eğer hipotez 2 kabul edilirse ikili bir eşdizim oluşturur. p , p_1 ve p_2 için maksimum olabilirlik 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 ve w_1w_2 ikilisinin gözlenme miktarlarıdır; N ise derleme ait toplam kelime sayısıdır. Binom dağılım kabulüyle $(b(k; n, x) = x^k(1-x)^{n-k})$ log-olabilirlik oranı, λ , şu şekilde tanımlanır:

$$\log \lambda = \log \frac{L(H1)}{L(H2)} = \log \frac{b(c_{12}, c_1, p)b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)b(c_2 - c_{12}, N - c_1, p_2)} \quad (4.2)$$

$$= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2 - c_{12}, N - c_1, p_2)$$

Bu tanımda $L(k, n, x) = x^k(1-x)^{n-k}$ olarak kabul edilir. Mood (1974) $-2\log \lambda$ dağılımının asimptotik olarak χ^2 dağılımı olduğunu göstermiştir (Dunning, 1993). Bu sebeple hesaplanan $-2\log \lambda$ değeri verilen önem aralığında χ^2 değerinden küçük ise bağımsızlığı simgeleyen hipotez; aksi durumda ise w_1w_2 ikilisinin bir eşdizim olduğunu belirten hipotez kabul edilir. Bu sebeple log-olabilirlik değeri bir hipotezin diğerine oranla ne kadar kabul edilebilir olduğunu gösteren bir değerdir. Bu değer arttıkça aday kelime ikilisi eşdizim olmaya o kadar yaklaşır. Yöntemin uygulanmasında derlemdeki tüm kelime ikililerinin log-olabilirlik değeri hasaplanır ve ikililer değer azalacak şekilde listelenir.

4.3.3.2. Pearson ki-kare (χ^2) testi (Pearson's chi-square test)

Pearson χ^2 testi kelime birliklerinin beklenen ve gözlenen sıklık değerlerine bağlı olarak eşdizimliliklerinin değerlendirildiği bir testtir. Bu testte, kelime birliğinde yer alan kelimelerin sıklık değerleri Çizelge 4.1'de verildiği üzere 2x2'lik bir tabloya yerleştirilir. Ki-kare testi için sıfır hipotezi “beyaz” ve “saray” kelimelerinin bağımsız olduğudur ve beklenen sıklık değeri ile ifade edilir. “beyaz saray” ikilisi için ölçülen yani gözlenen ortalama değer beklenen değerden farklılaştıkça ikilinin eşdizim olma ihtimali yükselir.

Çizelge 4.1. “beyaz” ve “saray” kelimeleri için gözlenen sıklık değerlerini içeren 2x2'lik tablo

	w1=beyaz	w1≠beyaz
w2=saray	8 (beyaz saray)	4667 (örneğin, kervan saray)
w2≠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.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (4.3)$$

Denklem 4.3’de i tablodaki satır, j ise sütun indeksini simgeler. Tablodaki her bir hücre için beklenen değer ilgili hücrenin satır toplamının sütun toplamı ile çarpılıp tablo toplamına bölünmesiyle hesaplanır. Derlem içindeki tüm kelime birlikleri için χ^2 değeri hesaplanarak azalan sırada listelenir. χ^2 değeri yüksek olan adaylar eşdizim olmaya en yatkın adaylar olarak kabul edilir.

4.3.3.3. T-testi (The t-test)

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

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \quad (4.4)$$

Denklem 4.4’de $\bar{x}$ örnek ortalama değeri (gözlenen ortalama değeri), s^2 varyans, N örnek büyüklüğü (derlemdeki toplam kelime sayısı) ve μ ise dağılımın ortalamasıdır.

w_1 ve w_2 kelimelerinin eşdizimliliklerinin tespitinde $f(w_1)$, $f(w_2)$ ve $f(w_1w_2)$ sırasıyla w_1 , w_2 ve w_1w_2 ‘nin sıklık değerini, N derlemdeki toplam kelime sayısını gösterirken, kelimelere ve ikiliye ait gözlenme olasılık değerleri şu şekilde tanımlanır:

$$\begin{aligned} P(w_1) &= \frac{f(w_1)}{N} \\ P(w_2) &= \frac{f(w_2)}{N} \\ P(w_1w_2) &= \frac{f(w_1w_2)}{N} \end{aligned}$$

$P(w_1w_2) = P(w_1) \cdot P(w_2)$ eşitliği sıfır hipotezini yani kelimelerin bağımsızlığını ifade eder. Eğer hipotez doğru ise rastgele bu ikilinin oluşturulması; bu sonuca 1 diğer tüm sonuçlara 0 değeri atama işlemi ortalama değeri $\mu = P(w_1) \cdot P(w_2)$ olan bir Bernoulli dağılımıdır. Binom dağılımı kullanarak örnek ortalaması $\bar{x} = P(w_1w_2)$ ve örnek varyansı $s^2 = P(w_1w_2)(1 - P(w_1w_2)) \approx P(w_1w_2)$ olarak hesaplanır ($P(w_1w_2)$ pek çok kelime ikilisi için düşük olacağından $1 - P(w_1w_2) \approx 1$ olarak kabul edilebilir). Bir kelime birliğinin t-testi sonucunun yüksek olması eşdizim ihtimalinin aynı derlemdeki diğer birliklere oranla daha yüksek olması anlamına gelir.

4.3.4. Ortalama –varyans yöntemi

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

Yöntem tüm eşdizimlerin ve kelimeler arası uzaklıkların listelenmesini gerektirir. Bir eşdizim adayı için verilen pencerede farklı uzaklıklarda görülme değerlerinin ortalama ve varyansı aşağıda verilen eşitlikler ile hesaplanır (Manning and Schütze, 1999) .

$$\bar{d} = \frac{\sum_{i=1}^{i=p} d_i}{p} \quad (4.5)$$

$$s^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1} \quad (4.6)$$

Denklem 4.5’de d_i bir eşdizim adayında ikinci kelimenin uzaklığını, p değeri ise ilgili eşdizim adayının derlem içinde kaç kere gözlemlendiğini verir.

Denklem 4.6’da verilen n bir eşdizim adayının kaç kere derlemde gözlendiğini göstermektedir.

Örneğin “kafa ye” birliğinin rastlandığı cümleler şu şekilde ise

- İşte kafayı şimdi yiyeceğim
- Bu problem bana kafayı yedirtti
- Kafayı yavaş yavaş yedi
- Kafayı yemeden bir atlatsak çok rahatlayacağız

İlk cümle için uzaklık değeri 2, ikinci ve dördüncü için 1, üçüncü için ise 3 olmaktadır. Bu durumda ortalama değer

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

şeklinde hesaplanır. Bu değer “kafa ye” birliği için ikinci kelimenin birinci kelimedenden ortalama olarak 1.75 kelime uzaklıkta yer aldığını gösterir. Varyans ise

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

şeklinde hesaplanır. Varyans değeri iki kelime arasındaki uzaklığın ne kadar değişken olduğunu gösterir. Değerin düşük olması iki kelimenin birbirinden sabit uzaklıklarda görülebileceğini dolayısıyla kelimelerin rastgele bir dizilişi olmadığını, değerın sıfır olması ise kelimelerin hep aynı uzaklıkta yer aldıklarının göstergesidir.

4.3.5. Smadja yöntemi

Smadja (1993) Xtract isimli çalışmasında eşdizimlerin tespiti için 3 temel özelliği göz önünde tutmuştur. Bunlar, eşdizim gücü, dağılımı ve kelimeler arası uzaklık değerleridir. Yöntem, bir kelimenin kendinden sonra gelen tüm kelimelerle ilişkisinin belirlenmesi vasıtasıyla hangi kelimenin bu kelimeyle bir eşdizim oluşturacağını tespitini içerir.

Eşdizim gücü, bir ikilinin eşdizim olma olasılığının aynı ilk kelimeyi içeren diğer ikililer vasıtasıyla hesaplanmış değeridir. Denklem 4.7 herhangi bir i kelime ikilisi için güç bilgisini vermektedir. $freq_i$ ilgili ikilinin sıklık değerini, $\bar{f}$ ortalama sıklık değerini, σ ise sıklık standart sapma değerini vermektedir.

$$güç = k_i = \frac{freq_i - \bar{f}}{\sigma}. \quad (4.7)$$

Güç ölçümünün yapılabilmesi için her kelime ve kendinden sonra gelen kelimeler bir pencere dâhilinde belirlenir ve sıklık değerleri ölçülür. Çizelge 4.2’de örnek olarak “maliye” kelimesi verilmiştir (değerler herhangi bir derlemde alınmamıştır). Öncelikli olarak aday kelimeye ait derlemde 5 pencere büyüklüğünde rastlanan tüm ikililer listelenmiş ve bu ikililerin sıklık değerleri ölçülmüştür.

“maliye” kelimesi ile oluşturulacak bir birlik için ortalama sıklık değeri

$$\frac{(10 + 7 + 4 + 2 + 2 + 1)}{6} = 4.33$$

şeklinde hesaplanır. Standart sapma değeri ise

$$\sqrt{\frac{(10 - 4.33)^2 + (7 - 4.33)^2 + (4 - 4.33)^2 + (2 - 4.33)^2 + (2 - 4.33)^2 + (1 - 4.33)^2}{6}}$$

olmaktadır. Daha sonra ise denklem 4.7 kullanılarak her bir eşdizim adayının gücü belirlenir.

Çizelge 4.2. Smadja yöntemi, örnek ikililer

birlik no	kelime1	kelime2	kelime birlik sıklığı	uzaklık				
				1	2	3	4	5
1	maliye	bakan	10	7	2	1	0	0
2	maliye	türkiye	7	1	2	2	1	0
3	maliye	üzeri	4	1	0	3	0	0
4	maliye	ziyaret	2	0	0	1	1	0
5	maliye	yarat	2	0	0	0	1	1
6	maliye	yap	1	1	0	0	0	0

Dağılım bilgisi her ikilide kelimeler arası uzaklık değerlerinden elde edilir. (denklem 4.8). Dağılım değerinin yüksek olması ikilinin sabit bir dizilişte yer aldığı sonucunu doğurur, yani uzaklık değerleri için çizilen histogram uç noktalara sahiptir.

$$dagi\ lim = U_i = \frac{\sum_{j=1}^5 (p_{i,j} - \bar{p}_i)^2}{5} \geq U_0 \quad (4.8)$$

Çizelge 4.3.5.1’de yer alan “maliye bakan” ikilisi için dağılım değeri $\bar{p}_1 = (7 + 2 + 1) / 5 = 2$

$$\sqrt{\frac{(7 - \bar{p}_1)^2 + (2 - \bar{p}_1)^2 + (1 - \bar{p}_1)^2}{5}} = \sqrt{\frac{(7 - 2)^2 + (2 - 2)^2 + (1 - 2)^2}{5}} = \sqrt{\frac{26}{5}}$$

şeklinde hesaplanır.

Dağılım ve güç kriterleri hangi kelime ikilisinin eşdizim olarak seçileceğini saptarken uzaklık bilgisi eşdizim içindeki iki kelimenin uzaklıklarını denklem 4.9 ile belirler. Denklemde k_1 , bir sabit olarak bulunmaktadır.

$$p_{i,j} \geq \bar{p}_i + (k_1 \times \sqrt{U_i}). \quad (4.9)$$

Smadja çalışmasında güç, dağılım ve uzaklık ölçümleri için (k_0, k_1, U_0) sayıları (1,1,10) olarak vermiştir.

4.3.6. Shimohata - Sugio - Nagata çalışması

Shimohata-Sugio-Nagata çalışması farklı uzunluktaki eşdizimlerin saptanmasına yönelik olarak geliştirilmiştir (Shimohata et al., 1997). Temel fikri, bir kelimenin eşdizim oluşturma ihtimalinin komşu kelimelerdeki değişkenlik değerine bağlı olduğudur. Eğer bir kelime dizisine ait komşu kelimeler çok değişkenlik gösteriyorsa bu dizi eşdizim olarak kabul edilir, aksi durumda komşu kelimelerle birleşerek eşdizim yaratması söz konusu olur. Yöntem iki basamakta özetlenebilir:

Basamak 1: Eşdizim birimlerinin özütlemesi

Eşdizim adaylarının belirlenmesi için kelime dizilerinin ardıl ve öncül kelimelerine ait dağılım bilgisine başvurulmaktadır. Bir kelime dizisini takip eden veya dizinin takip ettiği kelimeler ve sahip oldukları sıklık (frekans) bilgisi saptanarak ilgili kelime dizisinin entropi bilgisi belirlenir. Buradaki temel düşünce şu şekilde açıklanabilir: Bir kelime dizisi eş dizim özelliğine sahipse bu durumda dil içinde anlam taşıyan tek bir birim gibi davranacaktır. Bu sebeple kendinden önce ve sonra görülecek olan farklı kelime sayısı da yükselecektir. Oysaki eğer bir kelime dizisi eşdizim özelliğini tamamlayamamışsa bu durumda kendinden önce veya sonra göreceği kelime sayısı azdır ve bu kelimeler ile eşdizim oluşturma olasılığı yüksektir.

Yöntemde entropi bilgisi bir düzensizlik (İng. disorder) ölçüsü olarak sunulmuştur. Bir kelime dizisinin düzensizlik bilgisinin yüksek olması kendinden önce veya sonra çok sayıda farklı kelime görülmesi durumudur, düşük olması ise kelime dizisinin eşdizimin bir parçası olduğu ihtimalini kuvvetlendiren bir durumdur. Kelime dizisine str ve frekans bilgisine $freq(str)$ dersek, her olası komşu kelimenin olasılığı aşağıdaki eşitlik ile saptanır.

$$p(w_i) = \frac{freq(w_i)}{freq(str)} \quad (4.10)$$

Bu durumda bu kelime dizisine ait olan entropi değeri dizinin öncül ve ardılı olmak üzere iki bölge için aşağıdaki şekilde hesaplanır.

$$H(str) = \sum -p(w_i) \log p(w_i) \quad (4.11)$$

Öncül ve ardıl entropi değerlerinden küçük olan değer kelime dizisinin entropi değeri olarak kabul edilir. Bir str kelime dizisinin eşdizim adayı olarak görülmesi ancak entropi değerinin bir eşik değerden yüksek olması ile sağlanır. Çalışmada bu eşik değeri 1 olarak alınmıştır.

Basamak 2 : Eşdizimlerin özütlenmesi

Eşdizimlerin çıkarsanması aşamasında 4 temel işlem yürütülmektedir.

1. $str_i (i = 1, 2, \dots, n)$ kelime dizisinden str_k aday eşdizimi (key string) alınır. Bu eşdizimin geçtiği cümleler derlem içinden çekilir. Bu cümleler içinde ilk basamakta entropi kıstasını geçmiş olan kelime dizileri cümleler içinde işaretlenir. Bu kelime dizileri ileriki kısımlarda str_i olarak adlandırılacaktır.
2. str_k ve str_i 'nin olası farklı kombinasyonlarına kaç adet rastlandığı belirlenir. Eğer frekans değeri belirli bir eşik değerini T_{freq} aşıyorsa str_i değeri çıkarsanır artık bu kelime dizisi aday eş dizim ile birleşebilecek bir kelime dizisi durumundadır. Eşik değeri olarak çalışmada $T_{freq} = 0.1$ alınmıştır.
3. Her str_i ve str_j kelime dizisi incelenir ve sırasıyla aşağıdaki durumlar sınanır:
 - a. Eğer str_i ve str_j örtüşüyor veya bitişiyorsa ve aşağıdaki eşitsizliği sağlıyorsa bu diziler birleştirilir

$$\frac{freq(str_i, str_j)}{freq(str_i)} \geq T_{ratio} \quad (4.12)$$

- b. str_j kelime dizisi str_i kelime dizisini içeriyor ve aşağıdaki eşitsizliği sağlıyorsa str_i dizisi elenir.

$$\frac{freq(str_j)}{freq(str_i)} \geq T_{ratio} \quad (4.13)$$

T_{ratio} çalışmada 0.8 olarak alınmıştır.

Örnek olarak aşağıda verilen cümlelerde altı çizili kelimeler ilk basamağı geçmiş olan birlikler ve “**refer to**” aday eşdizim olarak alınırsa:

Refer to the appropriate manual for instructions on ...

Refer to the manual for specific instructions.

Refer to the installation manual for specific instructions for...

Refer to the manual for specific instructions on...

$T_{freq} = 2$ için “*the, manual, for specific instructions, on*” kelime dizileri frekans kısıdını geçebilirler. $T_{ratio} = 0.75$ alındığında ise “*manual for specific instruction*” kelime dizisi oluşur. Sonuçta bu işlemten sonra “*the, manual for specific instructions, on*” kelime dizileri eşdizim özelliği göstererek “*refer to*” adayına eklenirler. Sonuçta çıkarsanan eşdizim şu şekildedir:

“Refer to the ... manual for specific instructions on”

Oluşan bu eşdizimde üç nokta ile gösterilen yerlere farklı kelimelerin gelmesi söz konusudur.

Shimohata ve diğerleri tarafından önerilen bu yöntemde derlem içindeki olası tüm uzunluktaki kelime dizileri aday olarak işleme tabi tutulmaktadır. Dolayısıyla uzay karmaşıklığı çok yüksek bir ayıklama işlemi ilk basamakta karşımıza çıkmaktadır. Benzer şekilde entropi ölçümü içinde yüksek sayıda işlem gerektiren bir arama yapılması gerekmektedir. Çalışma sonucunda ilk basamak için %88.1, ikinci basamak için %66.9 duyarlılık değerleri rapor edilmiştir. Eşdizim belirlenmesi üzerine yapılan diğer pek çok çalışmada da olduğu gibi eşdizim bilgisi etiketli derlemin yoksunluğu sebebiyle anma değeri ölçülememiştir. Çalışmanın diğer yöntemlerden diğer bir farkı da eşdizim olarak kabul ettiği adaylardır. Çalışma içinde örnekte de görüleceği üzere çok geniş ve/veya göreceli bir eşdizim kavramı söz konusudur.

4.4. Türkçe’de Eşdizim 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 kelime türetmek mümkündür. Bu üretkenlik, özellikle hesaplamalı dil bilim alanında yapılan çalışmaların zaman ve bellek karmaşıklığını çok yükseltmektedir. Uygulamalardaki karmaşıklığın yanı sıra bir metnin gövdelenmiş hali³ ile yüzeysel hali⁴ arasındaki farklılık sebebiyle herhangi biri temel alınarak geliştirilen bir model veya yöntem diğeri üzerinde uygulandığında farklı sonuçlar üretebilmektedir.

Türkçe üzerine yapılan çalışmaların pek çoğu eşdizim kavramının çeviri veya yabancı dil öğretiminde önemini tartışan veya belirli bir kelimenin eşdizimlilik özelliğini araştıran çalışmalardır (Sarıkas, 2006; Taşıgüzel, 1988; Özkan, 2007).

Oflazer vd. Türkçe’de çoklu kelime ifadelerinin (İng. multi-word expression) tespiti için kural tabanlı bir çözüm önermişlerdir (Oflazer et al., 2004). Bu sistem morfolojik olarak analiz edilmiş yani sözcük türleri işaretlenmiş ve kelime ekleri belirlenmiş bir derlem üzerinde çalışmaktadır. Oflazer vd. çoklu kelime ifadelerini 4 farklı grupta değerlendirmiştir: kelimelerin ek almadığı sabit ifadeler, kelimelerin bir kısmının ek almadığı bir kısmının alabildiği yarı sabit ifadeler, kelime tekrarları veya zıtlıkları içeren ifadeler, adlandırılmış varlıklar. Çalışmada önerilen sistem belirli morfolojik desenleri 1100 kural çerçevesinde değerlendirerek çoklu kelime ifadelerini metin içerisinde tespit etmektedir.

4.5. Eşdizim Eğilimi Yöntemi

Tez kapsamında geliştirilen “eşdizim eğilimi” yöntemi, temel olarak kelimeler arası anlam bütünlüğü fikrine dayanmaktadır. Eşdizimi oluşturan kelimeler arasında anlam bütünlüğü olması sebebiyle kelimelerin birbirini çağrıştırdığı kabul edilmektedir. Bu sebeple bir kelime ikilisinin eşdizim olup olmadığına karar vermek amacıyla kelimeler arası ve birliğin diğer kelimelerle olan ilişkisi değerlendirilmelidir. Yöntem şu şekilde ifade edilebilir:

³ Metin içerisindeki kelimeler çekim eklerinden ayrılmışlardır

⁴ Metin içerisindeki kelimeler hem çekim hem yapım eklerine sahiptirler.

Eğer bir kelime anlam bütünlüğünü tamamlamamışsa, eşdizim oluşturma eğilimi gösterecektir. Bu durumda, bu kelimenin ardındaki kelimenin en az çabayla tahmin edilebilmesi gerekmektedir. Bu tahmin, kelimeyi takip eden farklı kelime miktarı ile ilgilidir. Kelimeye ait sıklık bilgisinin, takip eden farklı kelime sayısına oranı “eşdizim eğilimi” olarak adlandırılabilir. Eşdizim eğilimi yüksek olan bir kelime başka bir kelime ile eşdizim oluşturma çabası içindedir. Aksi durumdaki kelime ise anlam bütünlüğünü tamamlamıştır. Benzer şekilde eşdizim eğilimi yüksek olan bir kelime ve ardındaki kelimedenden oluşan kelime ikilisinin de eşdizim eğilimi ölçülebilir. Bu ikilinin ardından gelen farklı kelime miktarı düşük ise, ikili anlam bütünlüğünü tamamlamak için başka kelimelere ihtiyaç duymaktadır. Eğer kelime ikilisi ardındaki farklı kelime miktarı yüksek ise ikili anlam bütünlüğünü tamamlamış yani iki kelimeli bir eşdizim oluşturmuştur.

$f(w_1)$, $n(w_1)$, $f(w_1w_2)$, $n(w_1w_2)$ sırasıyla w_1 kelimesine ait sıklık değeri, w_1 kelimesini takip eden farklı kelime sayısı, eşdizim adayı w_1w_2 ’a ait sıklık bilgisi ve adayı takip eden farklı kelime sayısını ifade ettiğini kabul edersek, w_1 kelimesine ait eşdizim eğilimi ve w_1w_2 aday ikilisine ait eşdizim eğilimi şu şekilde hesaplanabilir:

$$T(w_1) = \frac{f(w_1)}{n(w_1)} \quad \text{ve} \quad T(w_1w_2) = \frac{f(w_1w_2)}{n(w_1w_2)}$$

$T(w_1)$, w_1 kelimesi için eşdizim eğilimini ifade eder. Bu değer ne denli yüksekse kelimenin eşdizim oluşturma isteği o denli yüksektir. $T(w_1w_2)$ ise kelime ikilisine ait eşdizim eğilimini ifade eder, bu değer ne denli düşüksen, bu ikili o denli eşdizim olmaya yatkındır. Bu değer belirli bir eşik değerinden düşük ise ikili anlam bütünlüğü tamamlamış ve anlam bütünlüğünü tamamlamak için bir başka kelimeye ihtiyaç duymamaktadır denilebilir.

Kelimeler arası gruplaşma eğilimi ilk olarak Altmann (1988) tarafından tartışılmıştır. Altmann (1988) iki kelimenin bir arada gözlenme sıklığına belirli bir pencere dâhilinde bakarak iki kelime arasındaki anlamsal ilişkiyi ölçen bir yöntem önermiştir. Tuzzi vd. (2010) çalışmasında Altmann’ın geliştirdiği yöntemi kullanarak İtalya Cumhurbaşkanı’nın yılsonu konuşmalarında sıklıkla rastlanan

kelimeler arasındaki anlamsal ilişkileri ortaya koymuşlardır. Eşdizim eğilimi yöntemi her ne kadar anlamsal olarak ilişkili kelimeleri yakalamak amacını taşımasa da kelimeler arası bir bütünlük arayışı bu yöntemde de söz konusudur. Bu benzerlik dikkate alındığında Altmann (1988) tarafından önerilen yöntemin ardışık iki kelimedenden oluşan eşdizimlerin tespiti için geliştirilmiş bir varyasyonu olarak görülebilir.

Tez kapsamında sadece iki kelimedenden oluşan eşdizimlerin tespiti üzerinde durulmuştur. Önerilen yöntem literatürde yer alan çeşitli yöntemlerin ürettiği listelerin filtrelenmesi amacıyla kullanılacaktır. Yöntem tez kapsamında iki kelimeleri aday eşdizimleri içeren listeler üzerinde test edilmiştir. Ancak önerilen bu yöntem/filtre tekrarlamalı olarak uygulanarak daha fazla miktarda kelime içeren eşdizimlerin tespitine de olanak tanımaktadır.

4.6. Deney Derlemleri

Eşdizim tespit yöntemlerinin Türkçe üzerinden sınanması için çeşitli dilbilim ve bilgi çıkarsama çalışmalarında kullanılmış olan Bilkent ve OSTAD derlemleri kullanılmıştır.

Bilkent derlemi Bilkent Üniversitesi'nde hesaplamalı dil-bilim çalışmalarının sonucu otomatik işaretlenmiş bir derlemdir (Tür et al., 2003). Derlemin içeriği, hazırlandığı yıllarda çıkan gazete yazıları ve makalelerden oluşmaktadır (Dinçer, 2004). Bilkent derleminin morfolojik analizi sonlu bir makine ile otomatik olarak yapılmıştır. Derlemin hem yüzeysel formu hem de gövde formu temin edilebilmektedir. Cümle sonlarının belirlenmesi ve derlemin gövdelenmiş hali benzer şekilde ilgili makine ile sağlanmıştır (Tür et al., 2003). Derlem, Dinçer (2004) tarafından tekrar düzenlenmiştir, bir takım hatalar ayıklanmıştır; tez çalışmasında düzenlenmiş olan bu son hali yer almıştır.

OSTAD, ODTU-Sabancı derlemi, (Ofłazer et al., 2003; Atalay et al., 2003) Türkçe üzerine yapılan birçok doğal dil işleme çalışmasında kullanılmış olan ODTU derleminin bir alt kümesinden oluşturulmuştur. ODTU derlemi içinden çekilen bu bölümün morfolojik analizi elle yapılmıştır. Otomatik analize kıyasla

elle oluşturulan morfolojik analiz sonuçlarının daha güvenilir olduğu düşünülmektedir. Çizelge 4.3 OSTAD ve Bilkent derlemine ait bir takım istatistiksel bilgileri içermektedir.

Çizelge 4.3. OSTAD ve Bilkent derlemlerinin genel özellikleri

	OSTAD	BILKENT
Kelime sayısı (Derlem büyüklüğü)	46532	719665
Kelime ikilisi sayısı	40867	671397
Cümle sayısı	5665	48268

4.7. Deneysel Hazırlık

4.7.1. Temel veri kümesi

Eşdizimlerin tespitinde tüm eşdizimlerin işaretlendiği bir derlem kullanılması ideal çözümdür. Ancak derlem büyüklükleri nedeniyle genel yaklaşım derlemlerden temel bir veri kümesi çekilerek bu veri kümesinin işaretlenmesi ve yöntemlerin bu veri kümesi üzerinde uygulanmasıdır. Temel veri kümesi pek çok farklı yöntemle (rassal seçim, sözcük türü dikkate alınarak, eşdizim sözlüğünden yararlanarak) oluşturulabilir (Pearce, 2002; Evert and Krenn, 2001).

Bu çalışmada, literatürde rastlanan yöntemlerden farklı olarak bilgi geri getirim alanındaki uygulamalara benzer bir temel veri kümesi oluşturulmuştur. Öncelikle derlem içerisinde yan yana gözlenen bütün farklı kelime ikilileri cümle sonları da dikkate alınarak belirlenmiştir. Bu ikililer üzerinde seçilen bazı istatistiksel yöntemler (görünme sıklığı, karşılıklı enformasyon, t testi, ki-kare testi vb.) uygulanarak her bir yöntem için ikili listeleri oluşturulmuştur. Bu listelerden en iyi değerlere sahip ilk 200 ikili, tek bir listede (toplam 1000 adet ikili) birleştirilmiştir. Yeni liste içinde tekrar eden ikililer çıkarılmıştır. Bu yeni liste üzerinde elle eşdizim işaretlemesi yapılmış ve eşdizim işaretli temel veri kümesi oluşturulmuştur. Bu şekilde oluşturulan veri kümesi ile farklı yöntemlerin ürettiği eşdizim aday listeleri eşit koşullarda kıyaslanabilmektedir.

Eşdizim olarak üç veya daha az sıklıkla gözlenen ikililer dikkate alınmamıştır. Ayrıca kelime ikililerinin anlam bütünlüğü oluşturduğu kabulüyle

cümle sonları dikkate alınarak, bir cümlemin sonunda bulunan kelime ile ardındaki cümlemin ilk kelimesinin oluşturduğu ikililer ve tek bir kelimeden oluşan cümleler deney dışı bırakılmıştır.

4.7.2. Değerlendirme yöntemi

Eşdizim tespit yöntemleri anma ve duyarlılık ölçütleri ile değerlendirilmiştir. Eşdizim tespitinde anma değeri, bir yöntem ile temel veri kümesi içinden çekilen doğru eşdizimlerin temel veri kümesi içinde yer alan doğru eşdizimlere oranı olarak tanımlanabilir. Duyarlılık değeri ise bir yöntemin eşdizim olarak atadığı ikililer içindeki doğru eşdizimlerin çekilen ikililere oranı olarak ifade edilir. ε değerini temel veri kümesi içinden çekilen, δ değerini ise temel veri kümesi içinde yer alan doğru eşdizim kümesi olarak kabul edersek, anma r ve duyarlılık p değerleri şu şekilde verilir:

$$r = \frac{|\varepsilon \cap \delta|}{|\delta|} \quad p = \frac{|\varepsilon \cap \delta|}{|\varepsilon|}$$

Çalışmamızda duyarlılık ve anma değerlerinin sunulmasında Evert ve Krenn çalışmasında (2001) izlenen yöntem tercih edilmiştir. Bu yaklaşımda, herhangi bir yöntemin ürettiği bir liste için tek bir noktada, örneğin ilk 5 ikili (N=5) gibi, duyarlılık ve anma değeri vermek yerine, listenin farklı noktalarında (örneğin N=1, 5, 10, ...) değerler ölçülür. Bu yöntemle herhangi bir anda yani tek bir N değeri ile varılabilecek hatalı sonuçlar engellenmiş olur ve yöntemler ürettikleri listelerin belirli anlarında birbirleri ile kıyaslanabilir hale gelirler. Tez kapsamında temel veri kümesinin her bir anı yani her bir N değeri için (N=1, 2, 3, ... ,temel veri kümesi toplam ikili sayısı) anma ve duyarlılık değerleri belirlenerek grafikler halinde sunulmuştur.

4.8. Sonuçlar

Eşdizim tespit yöntemlerinin ve eşdizim eğilimi yöntemlerinin Türkçe üzerinden sınanması için çeşitli dilbilim ve bilgi çıkarsama çalışmalarında kullanılmış olan Bilkent ve OSTAD derlemleri kullanılmıştır. Bu sınama süreci üç aşamada gerçekleştirilmiştir.

Ön değerlendirme: Literatürde yer alan yöntemler deney derlemleri üzerinde tatbik edilerek, her bir yöntemin kendi içinde Türkçe üzerinde uygulanabilirliği ve amaca uygunluğu incelenmiştir.

Yöntemlerin kıyaslanması: Türkçe'nin yapısına ve amaca uygunluğu kabul edilen yöntemler bu aşamada birbirleri ile kıyaslanmıştır. Değerlendirme yöntemi olarak duyarlılık ve anma grafikleri kullanılmıştır. Bu kıyaslama ile yöntemlerin gövdelenmiş ve yüzeysel formdaki derlemler üzerindeki etkileri de incelenmektedir.

Eşdizim eğilimi yönteminin uygulanması: Son aşama olarak, belirlenen yöntemler üzerine önerilen eşdizim eğilimi yöntemi tatbik edilerek yöntemlerin ürettiği sonuçlardaki değişimler/gelişimler değerlendirilmiştir.

İzleyen alt bölümler deneysel çalışma aşamaları tanıtmaktadır.

4.8.1.Ön değerlendirme

Tez çalışmasında öncelikle literatürde yer alan çeşitli eşdizim yöntemleri Türkçe derlemler üzerinde uygulanmıştır. Her bir yöntemin ürettiği sıralı liste diğer yöntemlerden bağımsız olarak değerlendirilmiştir. Bu ön çalışma esnasında Türkçe eşdizimlerin sıklıkla az sayıda kelime içerdiği ve ardışık kelimelerden oluştuğu görülmüştür. Bu sebeple, eşdizimi oluşturan kelimeler arasına farklı kelimelerin girdiği ve daha çok iki kelimedenden uzun eşdizimler üzerinde başarılı sonuçlar üreten Shimohata-Sugio-Nagata yöntemi çalışma dışı bırakılmıştır.

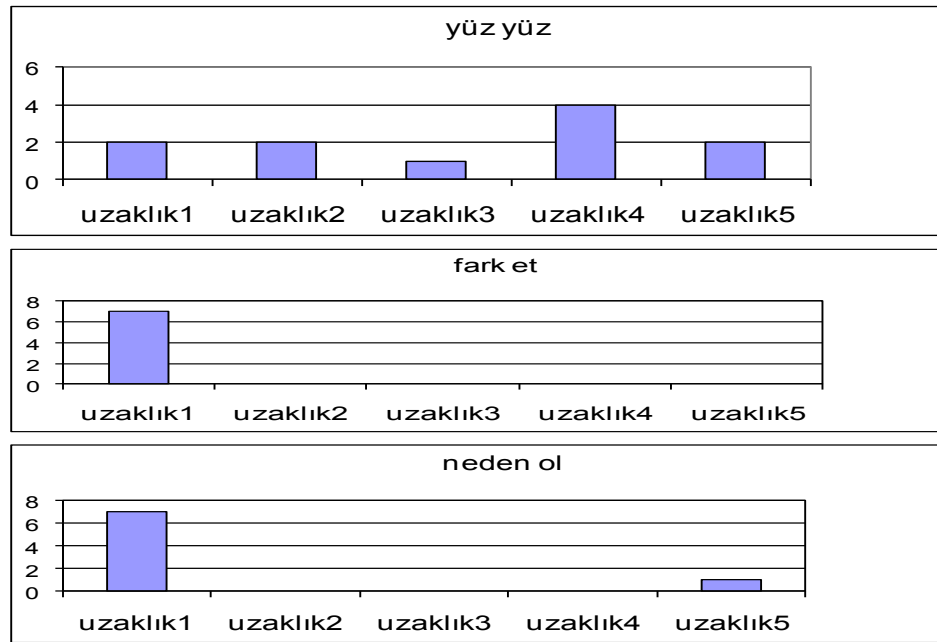
Ortalama-varyans yönteminin test edilmesi için öncelikle OSTAD derleminde örnek ikililer belirlenerek bu ikililer için yöntemin ürettiği sonuçlar gözlemlenmiştir. Derlemde pencere büyüklüğü=5 alınarak ilgili çalışma yapılmıştır (her kelime kendinden sonraki 5. kelimeye kadar eşleşmektedir). Çizelge 4.4'de “yüz yüz”, “fark et” ve “neden ol” ikilileri için sıklık ve uzaklık bilgileri verilmiştir. Uzaklık sütunlarında eşdizim adayı için ikinci kelimenin hangi uzaklıklarda kaç defa gözlemlendiği tutulmaktadır. Örneğin neden-ol ikilisi için

“ol” kelimesi “neden” kelimesinden 7 kez 1 uzaklıkta, 1 kez de 5 uzaklıkta görülmüştür.

Çizelge 4.4. OSTAD derleminden seçilen örnek eşdizim adayları için birarada görünme sıklık ve uzaklık değerleri

Eşdizim adayının gözlenme miktarı	kelime1	kelime2	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 4.1’de ise örnek eşdizim adaylarının uzaklıklarına ait histogramlar verilmiştir. Histogramlarda belli uzaklık değerlerinde yığılma görülen adaylar kelime birliği olarak kabul edilir. Eğer iki kelimenin tüm uzaklık değerlerinde görülme olasılığı eşit ise bu grubun eşdizim olmadığı kararına varılır.



Şekil 4.1. OSTAD derlemi örnek aday eşdizimlerine ait uzaklıkların grafiksel gösterimi

Çizelge 4.5 ve 4.6’da ortalama – varyans yönteminin ilk sonuçları OSTAD ve Bilkent derlemlerinin gövdelenmiş halleri için sıklık bilgisine göre azalan sırada listelenmiştir, listelerde ilk 60 ikili yer almaktadır. Listelerde standart sapma ($\sqrt{s^2}$) ve ortalama değerleri verilmiştir. Bu yöntemde, ortalama değeri 1 ve varyans (veya standart sapma) değeri düşük olan (sıfıra yakın) aday birlikleri ardışık kelimelerden oluşan kalıplaşmış sözcük gruplarını, ortalama değeri 1’den

çok yüksek ve varyans değeri düşük olan aday eşdizimler ise aralarına kelime girebilen eşdizimleri göstermektedir. Örneğin, “çizgi kahraman” ikilisi için “çizgi roman kahramanı”, “çizgi film kahramanı” gibi desenler ortaya çıkabilmektedir.

Çizelge 4.5. OSTAD derlemi ortalama –varyans yöntemi

sıklık	kelime1	kelime2	ortalama	standart sapma	sıklık	kelime1	kelime2	ortalama	standart sapma
14	bilimsel	devrim	1.57	1.45	7	arı	kırlangıç	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	gerçeklik	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	ver	2.50	1.64
8	uçak	kaza	1.00	0.00	6	ban	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

Ortalama-varyans yönteminde kelime 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. Bu durumda yöntem istenilen duyarlılık ve anma değerlerini üretemeyecektir. Tez kapsamında bir derlem veya metindeki iki ardışık kelimedenden oluşan eşdizimlerin tespiti amaçlanmaktadır. Bu

sebeple ortalama-varyans yönteminin diğer yöntemler ile kıyaslanmasının uygun olmayacağına karar verilmiştir.

Çizelge 4.6. Bilkent derlemi ortalama –varyans yöntemi

sıklık	kelime1	kelime2	ortalama	standart sapma	sıklık	kelime1	kelime2	ortalama	standart sapma
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

Smadja yöntemi özellikle bir anahtar kelime için eşdizim oluşturup oluşturmadığı ve hangi kelime ile eşdizim oluşturduğu konularında bilgi vermektedir. Tez kapsamında amaç bir derlem veya metinde yer alan eşdizimlerin tespitidir. Bu sebeple OSTAD derleminde adaylar güç değeri azalacak şekilde listelenmiştir. Bu durumda başka kelimeler ile hiç rastlanmayan sadece bir arada gözlenen kelimeler güç değerleri sonsuz olduğu için liste başında yer almaktadırlar (Çizelge 4.7). Listenin kalanı Smadja eşik değerleri dikkate alınarak seçilen birliklerdir.

Tez kapsamında Smadja yönteminin bir veya birkaç anahtar kelime yerine tüm derlem için uygulanarak diğer yöntemler ile kıyaslanması söz konusudur. Bu durumda yöntem yüksek duyarlılık değerleri üretirken yüksek zaman karmaşıklığı sorununu da beraberinde getirmektedir. Bu durum göz önünde bulundurularak yöntem temel veri kümesi oluşturulması aşamasına dâhil edilmemiştir.

Çizelge 4.7. OSTAD derlemi Smadja yöntemi sonuçları

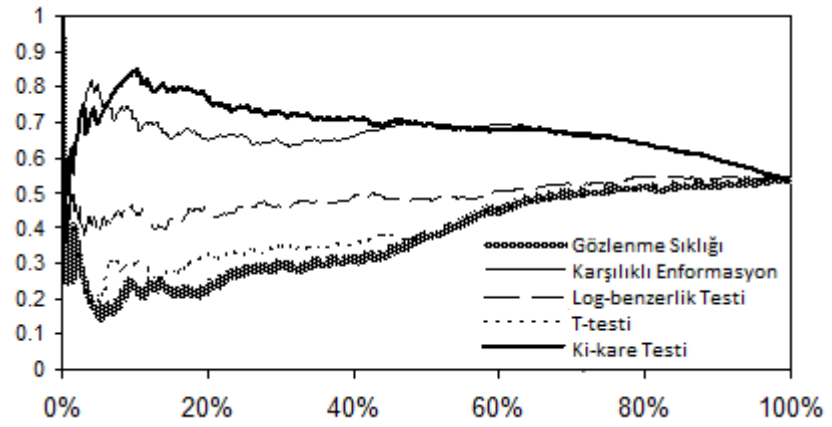
ikili sıklığı	kelime1	kelime2	güç	dağılım	uzaklık	ikili sıklığı	kelime1	kelime2	güç	dağılım	uzaklık
12	naci	bey	∞	23.04	1	6	devam	et	∞	5.76	1
14	bilimsel	devrim	∞	21.76	1	6	dikkat	çeken	∞	5.76	1
11	beyaz	peynir	∞	19.36	1	6	dikkat	çek	∞	5.76	1
11	karar	ver	∞	19.36	1	6	günah	çıkarma	∞	5.76	1
10	nusret	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	genelkurmay	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	uçaç	ilgili	1	1.2	3
6	çizgi	kahraman	∞	5.76	2						

Ön değerlendirme sonucunda gözlenme sıklığı, karşılıklı enformasyon, log-benzerlik, t testi ve ki-kare testinin en iyi sonuçları üretmeseler dahi uygulanabilirlik ve kıyaslanabilirlik konusunda diğer yöntemlere oranla daha avantajlı olduğu görülmüştür. Bu sebeple Bilkent derleminin hem yüzeysel hem gövde formundan elde edilen temel veri kümeleri üzerinde sadece bu yöntemlerin etkisi araştırılmıştır.

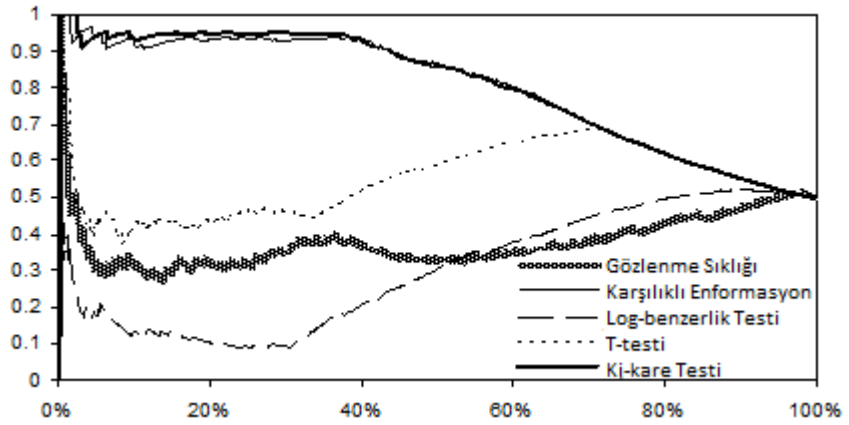
4.8.2. Yöntemlerin kıyaslanması

Ön değerlendirme sonucu seçilen eşdizim tespit yöntemlerinin; gözlenme sıklığı, karşılıklı enformasyon, log-benzerlik, t testi ve ki-kare testi; Bilkent derlemine uygulanmıştır. Her yöntem sıralı bir eşdizim aday listesi oluşturmaktadır. Bu listeler içinden her listenin ilk 200 aday seçilerek temel veri kümesi oluşturulmuştur. Gövdelenmiş derlemde (661 aday ikili) yüzeysel forma (506 aday ikili) oranla daha çok miktarda ikili içeren bir temel veri kümesi oluşmuştur. Bu veri kümeleri incelendiğinde gövde formu için %49.8, yüzeysel

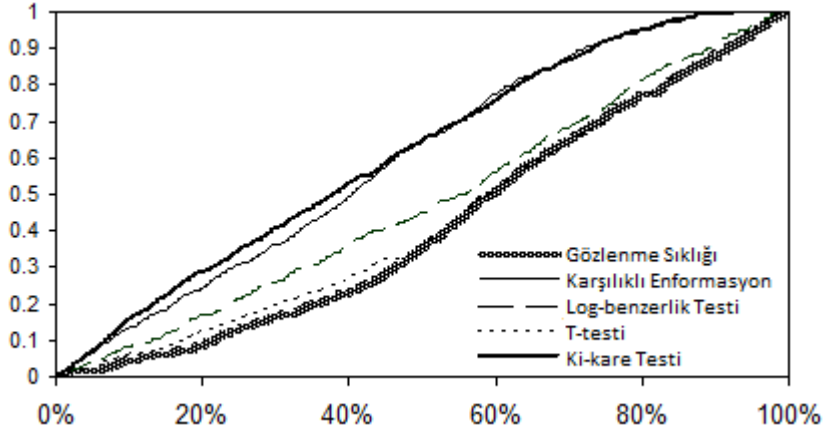
form için %53.5 oranında doğru eşdizime rastlanmaktadır. Bu değerler verilen temel veri kümesi üzerinde koşturulan yöntemler için taban değer (*baseline*) olarak kabul edilebilir. Şöyle ki tüm veri kümesi tamamlandığında yöntemlerin tümü için duyarlılık değeri bu taban değerine ulaşacaktır. Dolayısıyla bir yöntemin duyarlılık değerinin herhangi bir anda bu taban değerden daha düşük olması, temel veri kümesinde rastgele bir eşdizim kararı vermekten dahi daha başarısız olduğu anlamına gelecektir. Bilkent derleminin yüzeysel ve gövde formu için elde edilen duyarlılık ve anma değerleri Şekil 4.2, 4.3, 4.4 ve 4.5’de verilmiştir.



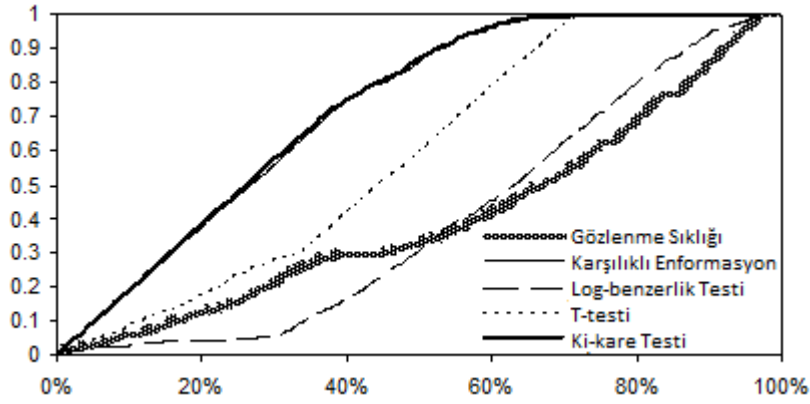
Şekil 4.2. Bilkent derlemi için duyarlılık grafiği (yüzeysel form).



Şekil 4.3. Bilkent derlemi için duyarlılık grafiği (gövde formu).



Şekil 4.4. Bilkent derlemi için anma grafiği (yüzeysel form).



Şekil 4.5. Bilkent derlemi için anma grafiği (gövde formu).

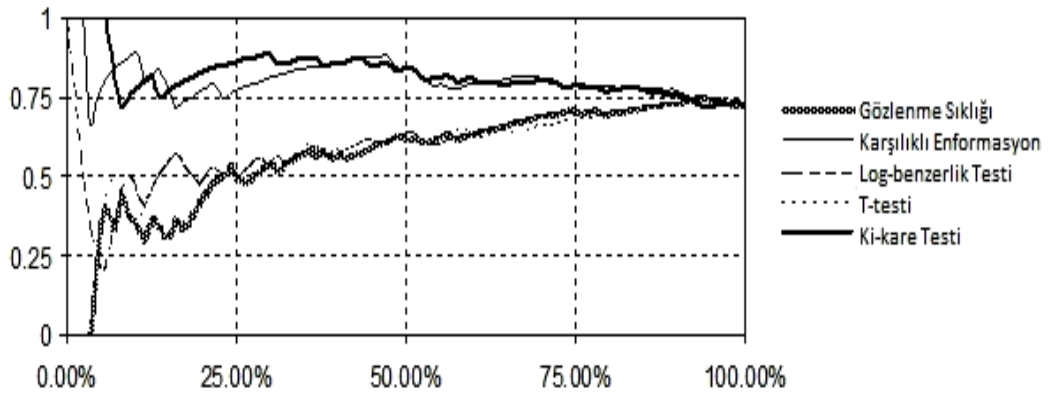
Şekil 4.2 ve 4.3 ile verilen duyarlılık grafiklerinde 3 temel özellik dikkat çekmektedir. İlk özellik χ^2 (ki-kare) ve karşılıklı enformasyon yöntemlerinin diğer yöntemlere oranla daha yüksek duyarlılık değerlerine sahip olmasıdır. Öyle ki log-olabilirlik, t-testi ve gözlenme sıklığı yöntemleri taban değer dahi altında kalmışlardır. Gözlenme sıklığı yöntemi diğer dillerdeki çalışmalarda da olduğu üzere tüm yöntemlere oranla daha başarısız olmuştur. İkinci özellik, gövde formu üzerinden çalıştırılan tüm yöntemlerin yüzeysel forma nazaran daha yüksek başarı göstermesidir. Duyarlılık grafikleri incelendiğinde ortaya çıkan üçüncü özellik, gövde formunun başarılı ve başarısız olarak gruplayabileceğimiz yöntemleri birbirinden ayırmada daha net sonuçlar ürettiğidir. En iyi yöntemler (χ^2 ve karşılıklı enformasyon) başarımlar açısından birbirlerine çok yaklaşırken diğer yöntemler başarımlar açısından arayı oldukça açmaktadırlar.

Duyarlılık grafiklerinin değerlendirmelerini destekleyen sonuçlar anma grafiklerinde de görülmektedir. Benzer şekilde χ^2 ve karşılıklı enformasyon yöntemleri doğru eşdizimleri tüm veri kümesi düşünüldüğünde kendi listeleri içinde daha üst sıralara yerleştirmekte dolayısıyla daha önce yakalamaktadır.

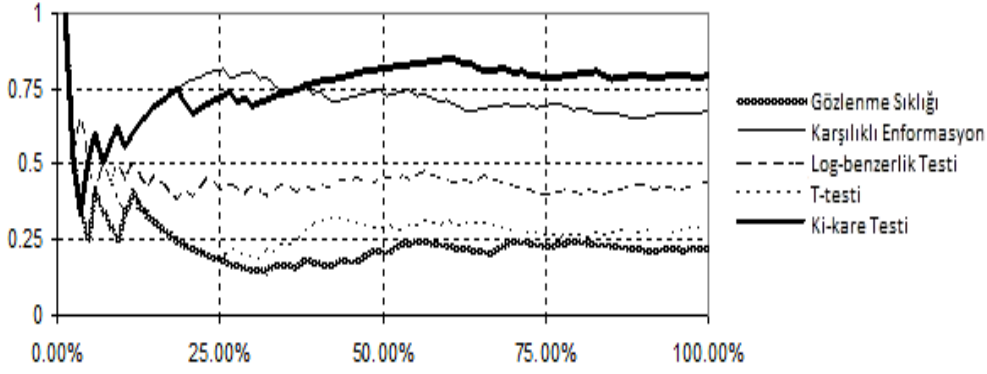
4.8.3. Eşdizim eğilimi yönteminin uygulanması ve değerlendirilmesi

Eşdizim eğilimi yönteminin sınanması için Bilkent derleminin yüzeysel formu kullanılmıştır. Gözlenme sıklığı, karşılıklı enformasyon, log-benzerlik, t testi ve kıkare testlerinin uygulanması sonucunda elde edilen temel veri kümesi üzerinde tüm küme için en yüksek duyarlılık değerini (0.724) üreten $(0.24)^{-1}$ değeri eşdizim eğilimi yöntemi için eşik değeri olarak belirlenmiştir. Yöntemin uygulanması sonucunda 506 aday ikilinin yer aldığı temel veri kümesi 87 ikili içeren bir liste haline gelmiştir.

Bu aday listesi her bir yönleme göre sıralanmış ve duyarlılık grafikleri hazırlanmıştır. Şekil 4.6 eşdizim eğilimi yöntemi sonrasında elde edilen duyarlılık eğrilerini içermektedir. Eşdizim eğilimi yönteminin sonuçlarının aynı sayıda aday içeren liste ile kıyaslanması amacıyla bu yöntem uygulanmadan önce diğer yöntemlerin ürettiği ilk 87 adaya ait duyarlılık grafiği de Şekil 4.7’de verilmiştir.



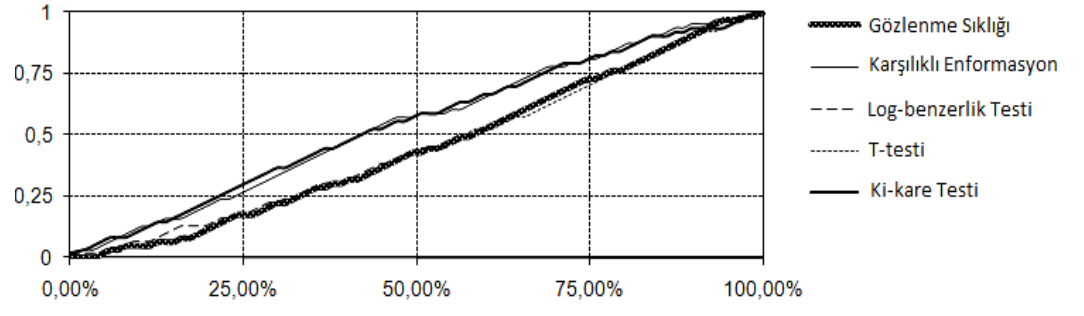
Şekil 4.6 Eşdizim eğilimi yönteminin uygulanması sonucu elde edilen duyarlılık grafiği



Şekil 4.7 Eşdizim eğilimi yöntemi öncesinde diğer yöntemlerin ürettiği ilk (en iyi) 87 aday ikili için duyarlılık grafiği

Şekil 4.6 ve 4.7’de yer alan duyarlılık grafikleri incelendiğinde eşdizim eğilimi yönteminin uygulanması sonucunda tüm yöntemlerde farklı oranlarda iyileşme görülmektedir. Öyleki yöntemlerin kendi başlarına ürettikleri listenin tümü değil sadece ilk 87 adayları dikkate alındığında (Şekil 4.7) dahi önerilen yöntemin duyarlılık değerlerine katkısı gözükmemektedir. Özellikle t testi ve kelime sıklığı yöntemleri diğer yöntemlere yaklaşacak oranda iyileşme göstermektedir. Bu sayede eşdizim tespiti için en basit yöntem olarak kabul edilen kelime sıklığı yöntemi diğer yöntemler ile yarışabilir hale gelmiştir. Ayrıca duyarlılık grafikleri incelendiğinde tüm yöntemlerin duyarlılık değerlerinin eşdizim eğilimi yöntemi ile birbirine yaklaştığı görülmektedir.

Şekil 4.8. eşdizim eğilimi yönteminin tatbiki sonucunda elde edilen anma eğrilerini içermektedir. Duyarlılık değerlerine benzer şekilde filtrelemeden sonra (Şekil 4.8) elde edilen anma değerleri ile filtrelemeden önce (Şekil 4.4) elde edilen anma eğrileri kıyaslandığında tüm yöntemlere ait anma eğrilerinin birbirine yaklaştığı, diğer bir deyişle yöntemlerin doğru eşdizimleri yakalama zamanlarının birbirine yaklaştığı görülmektedir.



Şekil 4.8. Eşdizim eğilimi yönteminin uygulanması sonucu elde edilen anma grafiği

5. SONUÇ

Bu tez kapsamında Türkçe metinlerin analizi metin içerisinde aktarılmak istenen enformasyonun Türkçe dilinin de özellikleri dikkati alınarak ortaya konmasını sağlayan tüm işlemler olarak tanımlanabilir. Bu işlemlerin temelinde, en küçük anlam birimi olan kelimeler arasında metnin anlam bütünlüğüne katkısı daha az olanların (işlev kelimeler) tespiti, taşıdığı anlamı güçlendirmek veya değiştirmek amacıyla başka kelime/kelimelerle birleşerek anlamsal bir bütün oluşturan kelimelerin (eşdizimler) belirlenmesi yer almaktadır. Bu tezde doğal dil işleme ve bilgi geri getirim gibi çeşitli çalışma alanlarına hizmet etmek üzere metinlerin anlamsal özelliklerini/birimlerini ortaya çıkarmayı hedefleyen bir çalışma sunulmuştur. Anlamsal katkı sağlamayan kelimelerin tespiti ve eşdizim olarak adlandırılan kelime birliklerinin hesaplamalı yöntemler ile tespiti tezin kapsamı dâhilindedir.

Metinler içinde aktarılmak istenen enformasyona katkısı az olan ya da hiç olmayan, temel görevi bağlaşık bağların kurulması olarak kabul edilen kelimeler metnin tümü dikkate alındığında azımsanamayacak kadar büyük bir yer işgal ederler. Bu kelimeler işlev kelimeler olarak isimlendirilmektedir. İşlev kelimelerin metinler içerisinde ayıklanması sayesinde metnin anlam bütünlüğünü oluşturan kelimelerin ortaya çıkarılması ve bu kelimeler ile metnin ifade edilebilmesi mümkün olur. İşlev kelimelerin tespitinde literatürde yer alan çalışmaların birçoğunda gözlenme sıklığı bilgisi kullanılmaktadır. Gözlenme sıklığı işlev kelimelerin tespitinde önemli bir özellik olmasına rağmen diğer pek çok unsurunda bu tespitte etkili olduğu düşünülmektedir. Bu sebeple farklı özelliklerin tek bir doğrusal fonksiyon vasıtasıyla bir arada değerlendirilebileceği istatistiksel bir yöntem olan diskriminant analizinin işlev kelimelerin tespiti için kullanılması önerilmiştir. Bu yöntem ve belirlenen özellikler farklı diller için de uygulanabilir olmakla beraber tez kapsamında yöntem Türkçe metinler üzerinde sınanmıştır.

Diskriminant analizi yöntemi iki temel basamaktan oluşmaktadır, talim ve sınama. Talim basamağında sözcük türleri etiketlenmiş bir derlem işlev kelime etiketli bir derlemin/metinlerin türetilmesi için kullanılmıştır. Talim derleminde

isim, fiil ve sıfat etiketli sözcükler dışındaki tüm kelimeler işlev kelime olarak kabul edilmiştir. Analize ait katsayılar ve eşik değer talim aşamasında çeşitli matris operasyonları ile elde edilmiştir. Analiz esnasında gerçekleşen matris operasyonlarının işlem karmaşıklığı yüksek olmasına rağmen bu işlemler sadece bir kere yürütülerek eşik değer ve katsayılar hesaplandıktan sonra sınama aşamasında farklı uzunluktaki farklı metinler için sayısız kere kullanılabilir.

Tez kapsamında tüm dillerdeki genel kurallar ve özel olarak Türkçe'nin dilbilgisel özelliklerinden yola çıkarak işlev ve içerik kelimeler arasındaki ayrımı sağlayan bağımsız değişkenler kelimenin cümle içindeki sıra numarası, kelime uzunluğu, kelime gözlenme sıklığı, eşdizimsel sıklık ve belge sıklığı olarak belirlenmiştir. Deneysel sonuçlar, gözlenme sıklığının en baskın özellik olmasına rağmen diğer özelliklerin analiz dışında bırakılması için yeterli bir kanıt olmadığını ortaya koymuştur.

Diskriminant analizi yönteminin deney basamağında, farklı konular içeren 8 metin sınama derlemi olarak kullanılmış; kelime gözlenme sıklığı, tf-idf ve diskriminant analizi yöntemleri için duyarlılık grafikleri verilmiştir. Gözlenme sıklığı ve tf-idf gibi sıralı liste üreten yöntemlerde işlev kelimeleri içerik kelimelerden ayıran bir eşik değer belirlenmesi eşik değerlerin başarımı doğrudan etkilemesi sebebiyle zor ve tartışmalı bir konudur. Bu çalışmada herhangi bir eşik değer belirlemek yerine listelerdeki ilk N aday işlev kelime olarak kabul edilmiştir. N değeri ise diskriminant analizi tarafından üretilen işlev kelime miktarını simgelemektedir. Gözlenme sıklığı ve tf-idf yöntemleri için aday listelerinde bu ilk N adayın ardında kalan kısımda duyarlılık değerinin hızla düştüğü görülmüştür.

Değerlendirme sonuçlarında diskriminant analizinin işlev kelimeleri içerik kelimelerden ayırt etmede çalışma içine dâhil edilen diğer yöntemlere oranla daha başarılı olduğu görülmüştür. Yöntem her sınama belgesi için ayrı bir işlev kelime listesi oluşturmaktadır. Ayrıca önerilen yöntem herhangi bir aday kelime için işlevlik kararı verirken derlem veya metin içinde yer alan diğer tüm kelimeler ile kıyaslama ihtiyacı da duymamaktadır. Böylece sıralı liste üreten diğer

yöntemlerdeki ön işlemlerin tümünü geride bırakarak işlevlik konusunda hızlı bir karar sunmaktadır.

Doğal dil içinde kimi kelimeler anlamlarını değiştirmek, güçlendirmek gibi çeşitli sebeplerle diğer kelimeler ile bir arada kullanılarak eşdizimleri oluştururlar. Eşdizimler hem anlamsal hem biçimsel olarak bir bütün oluşturdukları için eşdizim içinde yer alan kelimelerin birbirinden bağımsız olarak değerlendirilmesi doğru değildir. Bu sebeple en genel tabiriyle bir metindeki anlam birimlerinin ortaya çıkarılmasını amaçlayan bu tezde eşdizimlerin tespiti çalışması kapsam içine alınmıştır. Mevcut çalışmalarda sunulan ve yaygın olarak kabul gören çeşitli yöntemler (gözlenme sıklığı, karşılıklı enformasyon, hipotez testleri gibi) Türkçe derlemeler üzerinde sınanmış, tezin amacına ve Türkçe'nin yapısına uygun olanlar belirlenmiş; uygulama sonuçları verilmiştir. Uygulama sonuçlarının sunulmasında temel veri kümesi olarak adlandırılan bir veri kümesi hazırlanarak farklı yöntemlerin eşit şartlar altında kıyaslanması sağlanmıştır. Değerlendirme aşamasında duyarlılık ve anma ölçütleri kullanılmıştır. Türkçe derlemin gövde ve yüzeysel formunda yöntemler uygulanmıştır. Her iki derlem için de en başarılı yöntemler karşılıklı enformasyon ve ki-kare testi olmuştur. Uygulanan tüm yöntemler kelimelerin gövde formlarının yer aldığı derlemde yüzeysel forma nazaran daha yüksek başarı göstermiştir. Ayrıca gövde formu başarılı ve başarısız olarak gruplayabileceğimiz yöntemleri birbirinden ayırmada daha net sonuçlar üretmiştir. İleriki çalışmalarda ilgili yöntemin farklı dillerdeki derlemlere tatbiki, yöntemin hesaplamalı dilbilim içindeki farklı alanlarda kullanılması hedeflenmektedir.

Eşdizimler içinde yer alan kelimeler birbirlerini anımsatırlar. Öyle ki eşdizim içinde yer alan ilk kelime söylendiğinde anadili olarak o dili konuşan bir kişi ilk kelimenin ardından gelen kelimeyi vakit kaybetmeden tahmin edebilir. Bu fikirden yola çıkarak eşdizimlerin tespiti için tez kapsamında “eşdizim eğilimi” yöntemi geliştirilmiştir. Bu yöntemde bazı kelimelerin anlam miktarlarını arttırmak veya anlam bütünlüklerini tamamlamak amacıyla diğer kelimelere ihtiyaç duyduğu kabul edilerek bir kelimenin diğer bir kelime ile birleşme eğilimine “eşdizim eğilimi” adı verilmiştir. Eşdizim eğilimi yüksek olan bir kelimenin ardındaki kelimenin en az çabayla tahmin edilebilmesi gerekmektedir.

Bu tahmin, kelimeyi takip eden farklı kelime miktarı ile ilgilidir. Dolayısıyla her bir kelime için kelimenin gözlenme miktarının ardından gelen kelime miktarına oranı eşdizim eğilimini vermektedir. İki kelimedenden oluşan eşdizimler için ilk kelimenin eşdizim eğilimi yüksek (ardından gelen farklı kelime sayısı göreceli olarak düşük), ikiliye ait eşdizim eğilimi düşük (eşdizim ardından gelen farklı kelime sayısı göreceli olarak yüksek) olmalıdır. Yöntem yinelemeli olarak uygulanarak farklı miktarda kelime içeren eşdizimler için de uygulanabilir niteliktedir. Mevcut yöntemlerin ürettiği listeler üzerinde bir filtre olarak eşdizim eğilimi yöntemi uygulanmıştır. Eşdizim eğilimi filtresinin tüm yöntemlerin başarımlarını yükselttiği gözlenmiştir. Özellikle gözlenme sıklığı gibi en az gayret sarf edilen yöntemlerin başarımlarını diğer yöntemlere yaklaştırması bakımında eşdizim tespiti problemine büyük katkı sağladığı gözlemlenmiştir.

Bu tez metin içindeki anlamlı birimlerin doğru bir şekilde tespitini sağlaması yönünden otomatik çeviri, bilgi geri getirme, otomatik özetleme, sözcük türü belirleme gibi metni anlam yönünden güçlendiren ve/veya metnin dilin kurallarına uygun olarak oluşmasını sağlayan birimlerin tespitini gerektiren pek çok çalışmaya destek vermektedir. İleriki çalışmalarda önerilen yöntemin farklı dillerde, paralel derlemlerde, yinelemeli olarak uygulanması planlanmıştır.

KAYNAKLAR DİZİNİ

- Alpkoçak, A. ve Özkarahan, E.,** 1995, Bilgi Bulma Sistemleri için Otomatik Dizinleme Yöntemi, Bilişim '95, Dokuz Eylül üniversitesi, İzmir, 247-253.
- Adalar, D.,** 2005, Arapça ve Türkçe Ders Kitaplarındaki "Yardımlaşma" ve "Arkadaşlık" Konulu Metinlerin Karşılaştırılması: Bir Eşdizimsel Örüntüleme Çözümlemesi Örneği, Tömer Dil Dergisi, 129:63-84.
- Aitchison, J.,** 1991, Language change : Progress or decay?, 2nd. Edition, Cambridge: University Press, 20p.
- Aktaş, Ş.,** 2009, Edebi Metin ve Özellikleri Üzerine, A.Ü. Türkiyat Araştırmaları Enstitüsü Dergisi, 39:187-200.
- Altmann, G. ,** 1988, Wiederholungen in Texten, Bochum, Brockmeyer, 228p.
- Atalay, N.B., Oflazer, K. and Say, B.,** 2003, The Annotation Process in the Turkish Treebank, in Proceedings of the EACL Workshop on Linguistically Interpreted Corpora - LINC, Budapest, Hungary, 33-38.
- Athoğlu, Y.,** 1994, Yazılı ve Sözlü İletişim, Gazi Üniversitesi Eğitim Fakültesi Dergisi, 2(2): 247-258.
- Baayen, R.H.,** 1996, The effect of lexical specialization on the growth curve of the vocabulary, Computational Linguistics, 22: 455-480.
- Benson, M.,** 1990, Collocations and General-purpose Dictionaries, International Journal of Lexicography, 3 (1): 23-34.
- Bichakjian, Bernard H.,** 1988, Evolution in Language, Ann Arbor, MI, Karoma, 178p.
- Bisht, R. K., Dhami H.S. and Neeraj Tiwari, N. ,** 2006, An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations, Journal of Quantitative Linguistics, 13: 161-175.
- Blanc, J-M. and Dominey, P. F.,** 2004, Using prosodic information to discriminate between function and content words, In B. Bel & I. Marlien (Eds.), Speech Prosody, Lisbon: International Speech Communication Association, 4p.
- Bond F. and Shirai S.,** 2003, A Hybrid Rule and Example-based Method for Machine Translation, In M. Carl & A. Way (Eds.), Recent Advances in Example-based Translation, Dordrecht: Kluwer Academic Publishers, 49-54.

KAYNAKLAR DİZİNİ (devam)

- Can F., Koçberber, S., Balçık, E., Kaynak, C., Öcalan, H.Ç. and Vursavas, O.,** 2008, Information Retrieval on Turkish Texts, Journal of the American Society for Information Science and Technology 59 (3): 407-421.
- Christiansen, M. H. and Dale, R.,** 2003, Language Evolution and Change, Handbook of brain theory and neural networks (2nd ed.), MA: MIT Press, Cambridge, 1290p.
- Church, K. W. and Hanks, P.,** 1990, Word Association Norms, Mutual Information, and Lexicography, Computational Linguistics, 16(1):22-29.
- Davies, W. V.,** 1997, Essay Writing in German, Manchester University Press, 96p.
- Dillon, M. and Federhart, P.,** 1982, The use of discriminant analysis to select content-bearing words. Journal of the American Society for Information Science and Technology 33(4):245-253.
- Dinçer, B.T. and Karaoğlu, B.,** 2004, Sentence Boundary Detection in Turkish, ADVIS 2004, 255-262.
- Dinçer, T.,** 2004, Türkçe için istatistiksel bir bilgi geri-getirim sistemi, Doktora Tezi, U.B.E., Ege Üniversitesi, İzmir, 371s.
- Diri, B. and Amasyalı, M.F.** 2003, Automatic Author Detection for Turkish Text , ICANN/ICONIP-13th International Conference on Artificial Neural Network and 10th International Conference on Neural Information Processing, Istanbul, 138-141.
- Dolamic L. and Savoy J.,** 2010, When stopword lists make the difference, Journal of the American Society for Information Science and Technology, 61(1):200-203.
- Dunning, T.,** 1993, Accurate methods for the statistics of surprise and coincidence, Computational Linguistics, 19(1):61-74.
- Ergin, M.,** 2000, Türk Dil Bilgisi, Bayrak Yayınları, 408s.
- Evert, S. and Krenn, B.,** 2001, Methods for the qualitative evaluation of lexical association measures, Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Toulouse, France, 188-195.
- Filizok, R.,** 2001, Anlam Analizine Giriş, Ege Üniversitesi Yayınları, İzmir, 160s.
- Firth, J.R.,** 1957, Modes of Meaning, Papers in Linguistics 1934-51, Oxford University Press, 190-215.

KAYNAKLAR DİZİNİ (devam)

- Foskett, A.C.**, 1996, The Subject Approach to Information, 5th ed., Library Association, London, 574p.
- Haas, W.**, 1966, Linguistic Relevance. In Memory of J.R. Firth (ed. Bazell, C.E. et al), Longman, London, 116-148.
- Halliday, M.A.K.**, 1961, Categories of the Theory of Grammar, Word, 17(3): 241-292.
- Halliday, M.A.K. and Hasan R.**, 1976, Cohesion in English. Longman Group UK Limited, New York, 392p.
- Hao, L. and Hao, L.**, 2008, Automatic Identification of Stop Words in Chinese Text Classification, In H. Zhou (Ed.) Proceedings of the 2008 International Conference on Computer Science and Software Engineering (CSSE), Danvers : IEEE Computer Society Conference Publishing Services, 718-722.
- Hawkins, J. A. and Murray G. M.**, 1992, The evolution of human languages, Santa Fe Institute Studies in the Sciences of Complexity XI, AddisonWesley, Reading, 361p.
- Herdan, G.**, 1960, Type-token mathematics: A textbook of mathematical linguistics. 'S-Gravenhage, the Netherlands: Mouton, 448p.
- Hindle, D.**, 1990, Noun Classification from Predicate-Argument Structures, Annual Meeting of the Association for Computational Linguistics (ACL 1990), Pittsburgh, Pennsylvania, USA, 268-275.
- Hoey, M.**, 1991, Patterns of Lexis in Text, Oxford University Press, 276p.
- Jiang X., Fan, X. , Wang, Y. and Jia K.**, 2009, Improving the Performance of Text Categorization Using Automatic Summarization., In S. S. Mahmoud, K. Jusoff and K. Li (Eds.), International Conference on Computer Modeling and Simulation (ICCMS), Danvers: IEEE Computer Society Conference Publishing Services, 347-351.
- Justeson, J.S. and Katz, S.M.**, 1995, Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns, Computational Linguistics, 21(1):1-27.
- Kalaycı,Ş., Albayrak, A.S. ve Eroğlu, A.**, 2006, SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri, Asil Yayın Dağıtım, 426s.
- Keçik, İ. ve Uzun, L.**, 2003, Türkçe Sözlü ve Yazılı Anlatım, Eskişehir, 74s.

KAYNAKLAR DİZİNİ (devam)

- Kita, K., Kato, K., Omoto, T. and Yano, Y.,** 1994, A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria, *Journal of Natural Language Processing*,1(1): 21-33.
- Kornai, A.,** 2002, How many words are there?, *Glottometrics* 2002/4, 61-86.
- Krippendorff, K.,** 2004, *Content Analysis: An Introduction to Its Methodology*. 2nd edition, Thousand Oaks, CA: Sage, 440p.
- Kumova-Metin, S.,** 2008, Türkçe’de Kullanılan İşlev Kelimelerin Zipf 1. Kanunu Esasında Değerlendirilmesi, *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 23(2): 467-475.
- Lahtinen,T,** 2000, *Automatic Indexing: An Approach Using an Index Term Corpus and Combining Linguistic and Statistical Methods*, PhD. Thesis, University of Helsinki, 228p.
- Lass, R.,** 1980, *On explaining language change*, Cambridge: University Press, 186p.
- Lazarinis F.,** 2007, Engineering and utilizing a stopword list in Greek web retrieval, *Journal of the American Society for Information Science and Technology* 58(11):1645-1652.
- Leech, G.,** 1974, *Semantics*. London, Middlesex: Penguin, 386p.
- Lyons, J.,** 1966, Firth's Theory of Meaning, In *Memory of J.R. Firth* (ed. Bazell, C.E. et al), 15p.
- Mandelbrot, B.,** 1959, A note on a class of skew distribution function analysis and critique of a paper by H.A. Simon, *Inform. and Control*, 2:90-99.
- Mandelbrot, B.,** 1952, An information theory of the structure of the language, *Symposium on Applications of Communication Theory*, 486--502.
- Manning, C.D. and Schütze, H.,** 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, 718p.
- Manning, C.D., Raghavan, P. and Schütze, H.,** 2008, *Introduction to Information Retrieval*, Cambridge University Press, 581p.
- Martindale, C. and McKenzie, D.,** 1995, On the Utility of Content Analysis in Author Attribution: The Federalist, *Computers and the Humanities*, 29(4): 259-70.

KAYNAKLAR DİZİNİ (devam)

- Martínez-Santiago, F., García-Cumbreras M.A., Díaz-Galiano M. C., and Ureña, L. A.,** 2004, Using Machine Translation Resources with a Mixed 2-Step RSV Merging Algorithm, In C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, & B. Magnini (Eds.), Multilingual Information Access for Text, Speech and Images (SINAI at CLEF 2004), 3491/2005, 156-164.
- Mc Intosh, A. and Halliday, M.A.K.,** 1966, Patterns of Languages: Papers in General, Descriptive and Applied Linguistics, London: Longman, 199p.
- McMahon, A. M. S.,** 1994, Understanding language change, Cambridge University Press, 376p.
- Mikheev, A.,** 2000, Tagging Sentence Boundaries, ANLP 2000, Washington, USA, 264-271.
- Morrison, D.G.,** 1969, On Interpretation in Discriminant Analysis, Journal of Marketing Research 6(2):156-163.
- Newmark, P.,** 1978, Approaches to Translation, Pergamon Press, Oxford, 319s.
- Oflazer, K., Çetinoğlu, Ö. and Say, B.,** 2004, Integrating Morphology with Multiword Expression Processing in Turkish, 2nd ACL Workshop on Multiword Expressions: Integrating Processing (MWE-2004), Barcelona, Spain, 8p.
- Oflazer, K., Say, B., Hakkani-Tür, D. Z. and Tür, G.,** 2003, Building a Turkish Treebank, Invited chapter in Building and Exploiting Syntactically-annotated Corpora, Anne Abeille Editor, Kluwer Academic Publishers, 261–277.
- Overall, J.E. and Klett, C. J.,** 1972, Applied Multivariate Analysis. McGraw Hill Book Company, 500p.
- Özkan, B.,** 2007, Türkiye Türkçesinde Belirteçlerin Fiillerle Birliktelik Kullanımları ve Eşdizimliliği, Doktora Tezi, Çukurova Üniversitesi, Adana, 573s.
- Palmer, D.D. and M.A. Hearst,** 1997, Adaptive multilingual sentence boundary disambiguation. Computational Linguistics, 23(3): 241- 267.
- Pearce, D.,** 2002, A comparative evaluation of collocation extraction techniques. Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), 1530-1536.
- Powers, D.M.,** 1998, Applications and explanations of Zipf's laws, In: D. Powers (ed.), NEMLAP3/CONLL98, New methods in language processing and Computational natural language learning, 151-160.

KAYNAKLAR DİZİNİ (devam)

- Prasad, B.D., 2008**, Content analysis: A method of Social Science Research , In D. K. Lal Das (ed) Research Methods for Social Work, New Delhi: Rawat Publications, 173-193.
- Putnins, T. J., Signoriello, D. J., Jain, S., Berryman, M. and Abbott, D., 2005**, Advanced text authorship detection methods and their application to biblical texts, Proc. SPIE: Complex Systems 6039 ed. Axel Bender, Brisbane, Qld., Australia, 6039:163-175
- Radev, D. R. and Fan, W., 2000.**, Automatic summarization of search engine hit lists. Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 99-109.
- Reynar , J. C. and Ratnaparkhi, A., 1997**, A Maximum Entropy Approach to Identifying Sentence Boundaries, Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington, DC, 16-19.
- Riley, M. D., 1989**, Some applications of tree-based modelling to speech and language. In DARPA Speech and Language Technology Workshop, Cape Cod, Massachusetts, 339-352.
- Ritt, N., 1995**, Language change as evolution: looking for linguistics genes, Vienna English Working Papers 4(1): 43-57.
- Samuelsson, C., 1996**, Relating Turing's Formula and Zipf's Law, Proc. Forth Workshop on Very Large Corpora, 70-78.
- Sarıkaş, F., 2006**, Problems in Translating Collocations, Elektronik Sosyal Bilimler Dergisi, 5(17):33-40.
- Savoy, J., 1999**, A stemming procedure and stopword list for general French corpora, Journal of the American Society for Information Science and Technology 50(10):944-952.
- Shimohata S., Sugio T. and Nagata J., 1997**, Retrieving collocations by co-occurrences and word order constraints, The Eighth Conference on European Chapter of the Association for Computational Linguistics, Madrid, Spain, 476-481.
- Sinclair, J.M., 1991**, Corpus, Concordance, Collocation, Oxford University Press, Oxford, 170p.

KAYNAKLAR DİZİNİ (devam)

- Sinka M. P. and Corne, D. W.,** 2003, Towards modernised and Web-specific stoplists for Web document analysis, Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence, Washington :IEEE Computer Society, 396-402.
- Smadja, F.A.,** 1993, Retrieving Collocations from Text: Xtract, Computational Linguistics, 19(1):143-177.
- Stemler, S.,** 2001, An overview of content analysis. Practical Assessment, Research & Evaluation, 7(17):479-498.
- Taşgüzel, S.,** 1988, İlköğretim Türkçe Ders Kitaplarında Öğretici Nitelikli Metinlerdeki Eşdizimsel Örüntülerin Görünümü, Dil Dergisi, Ankara Üniversitesi Türkçe ve Yabancı Dil Araştırma ve Uygulama Merkezi, 25:72-87.
- Tsz-Wai Lo, R., He, B. and Ounis, I.,** 2005, Automatically building a stopword list for an information retrieval system, Journal Of Digital Information Management: Special Issue On The 5th Dutch-belgian Information Retrieval Workshop, 3(1):3-8.
- Tuzzi, A., Popescu, I.-I. and Altmann, G.,** 2010, Quantitative Analysis of Italian Texts. Lüdenscheid: RAM-Verlag, 5p.
- Tür, G., Hakkani-Tür, D. and Oflazer K.,** 2003, A Statistical Information Extraction System for Turkish, Natural Language Engineering, 9(2):181-210.
- Van Buren, P.,** 1967, Preliminary Aspects of Mechanisation in Lexis. Cahiers de Lexicology, 89-112.
- Van der Wouden, T.,** 1997, Negative Contexts: Collocation, Polarity and Multiple Negation, London, GBR, Routledge, 302p.
- Van Dijk, T.A.,** 1981, Studies in the Pragmatics of Discourse, Berlin/New York: Mouton Publishers, 331p.
- Van Rijsbergen, C.J.,** 1979, Information Retrieval (2nd ed.), Butterworths, London, 224p.
- Weber, R. P.,** 1990, Basic Content Analysis, 2nd ed. Newbury Park, CA, 96p.
- Witten, I. H., Moffat, A. and Bell, T C.,** 1999, Managing Gigabytes: Compressing and Indexing Documents and Images. San Francisco: Morgan Kaufmann, 519p.

KAYNAKLAR DİZİNİ (devam)

- Zhang, Y. and Wildemuth, B. M.**, 2009, Qualitative analysis of content, In B. Wildemuth (Ed.), Applications of Social Research Methods to Questions in Information and Library Science Westport, CT: Libraries Unlimited, 308-319.
- Zipf, G. K.**, 1949, Human Behaviour and the Principles of Least Effort, Cambridge, MA, Addison-Wesley, 573p.
- Zou, F., Wang, F.L., Deng, X. and Han, S.**, 2006, Automatic identification of Chinese stop words. Research on Computing Science 18: 718-722.

ÖZGEÇMİŞ

1979 yılında Ankara’da doğan Senem Kumova Metin, 1997 yılında Alanya Ayşe Melahat Anadolu Lisesi’ni bitirdikten sonra Ege Üniversitesi Elektrik Elektronik Mühendisliği Bölümü’nde lisans öğrenimine başlamış ve bu bölümden 2001 yılında mezun olmuştur. Yine 2001 yılında yüksek lisans öğrenimine başladığı Ege Üniversitesi Uluslararası Bilgisayar Enstitüsü’nden 2005 yılında Bilgi Teknolojileri Bilim Dalı’nda yüksek lisans derecesi ile mezun olmuş ve aynı yıl içerisinde bu enstitüde doktora öğrenimine başlamıştır. İzmir Ekonomi Üniversitesi, Mühendislik ve Bilgisayar Bilimleri Fakültesi, Yazılım Mühendisliği Bölümünde 2004-2005 yılları arasında araştırma görevlisi olarak çalışmıştır. 2005 yılından bugüne kadar ise aynı bölümde öğretim görevlisi olarak görev yapmaktadır.

Bugüne kadarki mesleki ilgi alanlarının hesaplamalı dilbilim, doğal dil işleme, derlem dilbilim olarak sıralanabileceği Senem Kumova Metin’in ilgili araştırma konularında çeşitli bilimsel dergilerde yayınlanan makaleleri ve çeşitli ulusal ve uluslararası bilimsel konferanslarda sunulan bildirileri bulunmaktadır.