TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

TÜRKÇE'DE EŞ ANLAMLI KELİMELERİN HESAPLAMALI DİYAKRONİK ANALİZİ

YÜKSEK LİSANS TEZİ

Umur Togay Yazar

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Prof. Dr. Ahmet Murat Özbayoğlu

ARALIK 2024

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Umur Togay Yazar

ÖZET

Yüksek Lisans

TÜRKÇE'DE EŞ ANLAMLI KELİMELERİN HESAPLAMALI DİYAKRONİK ANALİZİ

Umur Togay Yazar

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Ahmet Murat Özbayoğlu

Tarih: Aralık 2024

Türkçe 1920'lerden günümüze kadar gelen süreçte Dil Devrimi'nin etkisiyle önemli yapısal değişikliklere uğradı. Dil Devrimi'nin amaçlarına uygun olarak Türkçe üzerinde yabancı dillerin etkisini azaltacak şekilde adımlar atıldı. Bu müdahaleler sebebiyle dile pek çok yeni kelime girerken bir o kadar kelime de günümüzde kullanılmamaya başlandı. Kelime dağarcığındaki değişimlerin dışında dilin imla ve gramer kurallarında da değişiklikler görüldü. Bu çalışmada Türkçe'nin bu hızlı değişimini motivasyon kabul ederek bu çalışmada eş anlamlı kelimelerin diyakronik hesaplamalı diyakronik analizi gerçekleştirildi. İlk olarak 1920'lerden 2020'lere kadar dönemi kapsayan modern Türkçe'nin en büyük diyakronik korpusu *Turkronicles*'ı oluşturuldu ve bu korpus üzerinde frekans analizleri gerçekleştirildi. Ayrıca korpusla beraber diyakronik analizlerde n-gram'lar, diyakronik kelime vektörleri, bir yazılım kütüphanesi, dijitalleştirilmiş karşılıklar sözlüğü gibi çeşitli kaynakları kullanıma sunuldu. Analizlerimiz sonucunda Türkçe'nin kelime dağarcığının yıllar geçtikçe önemli ölçüde değiştiğini gördük. İmla kurallarının nasıl değiştiğini gözlemlemek için spesifik olarak sonu -b/-p ve -d/-t ile biten kelimeleri inceledik ve benzer bir sonuçla karşılaşıldı. Analizler bütünüyle ele alındığında oluşturduğumuz kaynakların aslında diyakronik analiz için uygunluğunu gösterdi ve

dilin ne ölçüde değiştiğini gözler önüne serdi. Ardından bu kaynakları kullanarak farklı tarihlerde kullanılmış eş anlamlı kelimeleri tespit etmeye yarayan iki farklı yöntem önerdik. Bu yöntemler esasında ortogonal dönüşüm matrisini bulmaya dayanır. Sonrasında önerdiğimiz yöntemler doğrusal dönüşüm matrisine dayalı temel bir yöntem ile karşılaştırıldı. Yöntemlerin başarısı iki farklı vektör algoritması üzerinden test edildi. Buna ek olarak zaman farkı arttıkça yöntemlerin performansının nasıl değiştiği gözlemlenip değerlendirildi. Deneylerin sonuçlarında önerilen yöntemlerin temel yöntemden daha iyi bir performans sergilediğini ve zaman farkına daha dayanıklı olduğu gözlemlendi. Son olarak eş anlamlı kelimelerin zamanla anlamının nasıl değiştiğini görmek adına çeşitli deneyler yürütüldü. Spesifik olarak eş anlamlı kelimelerin anlam daralması ve genişlemesi dinamikleri incelendi. İlk olarak *Turkronicles* korpusunda yer alan hatalı kelimeleri düzeltmek için bir Kodlayıcı-Çözücü modeli eğitildi ve bir test kümesi oluşturuldu. Hatalı kelimeler düzeltildikten sonra anlam daralması ve genişlemesini kelime ilişki çizgesi ve bir merkeziyet ölçütü etrafında modellendi. Ardından bu ölçütün bir eş anlamlı çifti için zaman boyunca nasıl değiştiğini gözlemlemek için Spearman korelasyon katsayısını hesaplandı. Son olarak birbirlerinin değişimini ne kadar etkilediğini ölçmek için bir Doğrusal Karma Model eğitildi. Deney sonuçlarına dayanarak eş anlamlı bir kelimenin anlamı genişlerken diğerinin anlamının daraldığı tespit edildi.

Anahtar Kelimeler: Diyakronik analiz, Doğal dil işleme, Türkçe korpus, Eş anlamlılar, Semantik değişim, Dil değişimi

ABSTRACT

Master of Science

COMPUTATIONAL DIACHRONIC ANALYSIS OF TURKISH SYNONYM WORDS

Umur Togay Yazar

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Computer Engineering Science Programme

Supervisor: Prof. Dr. Ahmet Murat Özbayoğlu

Date: December 2024

Turkish has undergone significant structural changes from the 1920s to the present day, influenced by the Turkish Language Reform. In line with the aims of the Language Reform, steps were taken to reduce the influence of foreign languages on Turkish. As a result of these interventions, many new words were introduced into the language, while many words fell out of use over time. Beyond changes in vocabulary, modifications were also observed in the spelling and grammar rules of the language. This study takes the rapid evolution of Turkish as its motivation and conducts a diachronic computational analysis of synonym words. First, we created *Turkronicles*, the largest diachronic corpus of modern Turkish, covering the period from the 1920s to the 2020s, and performed frequency analyses on this corpus. Additionally, we made various resources available for diachronic analyses, such as n-grams, diachronic word vectors, a software library, and a digitized dictionary of synonym words. Our analyses revealed that Turkish vocabulary has significantly changed over the years. To observe changes in spelling rules, we specifically examined words ending in -b/-p and -d/-t and encountered similar results. Overall, our analyses demonstrated the suitability of the resources we created for diachronic analysis and highlighted the extent of evolution of Turkish. Using these resources,

we proposed two different methods to identify synonym words used in different time periods. These methods are essentially based on finding an orthogonal transformation matrix. We then compared the proposed methods to a baseline method, which relies on a linear transformation matrix. The effectiveness of our methods was tested using two different vector algorithms. Furthermore, we observed and evaluated how the performance of the methods varied with increasing temporal differences. Our experimental results showed that the proposed methods outperformed the baseline method and were more robust to temporal differences. Finally, we conducted various experiments to explore how the meanings of synonymous words change over time. Specifically, we examined the dynamics of semantic specialization and generalization of synonymous words. Initially, we trained an encoder-decoder model to correct misspelled words in the *Turkronicles* corpus and created a test set. After correcting the misspelled words, we modeled semantic specialization and generalization using a word association graph and a centrality measure. To observe how this measure changes over time for a pair of synonyms, we calculated the Spearman correlation coefficient. Lastly, we trained a Linear Mixed Model to measure the extent to which the changes in one synonym affect the other. Based on the experimental results, we found that when the meaning of one word generalizes, the meaning of its synonym tends to specializes.

Keywords: Diachronic analysis, Natural language processing, Turkish corpora, Synonym words, Semantic change, Language change

.

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocalarım Mücahid Kutlu, Ahmet Murat Özbayoğlu, İsa Kerem Bayırlı ve Mehmet Akşit, kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine teşekkürlerimi ve minnetlerimi sunarım. Bu çalışmayı beni bu süreçte hiç yalnız bırakmayan hayat arkadaşıma ve aileme adıyorum.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

KISALTMALAR

ASR : Otomatik Konuşma Tanıma (Automatic Speech Recognition)
CBOW : Sürekli Kelime Torbası (Continuous Bag-of-Words)
CER : Karakter Hata Oranı (Character Error Rate)
DDİ : Doğal Dil İşleme
GB : Giga Byte
HTML : Hypertext Markup Language
JI : Jaccard Indeksi (Jaccard Index)
JSD : Jensen-Shannon Sapması (Jensen-Shannon Divergence)
JSON : JavaScript Obje Gösterimi (JavaScript Object Notation)
LDA : Gizli Dirichlet Dağılımı (Latent Dirichlet Allocation)
LLM : Büyük Dil Modeli (Large Language Model)
LM : Dil Modeli (Language Models)
LMM : Doğrusal Karma Model (Linear Mixed Models)
LT : Doğrusal Dönüşüm (Linear Transformation)
MRR : Ortalama Karşılıklı Sıra (Mean Reciprocal Rank)
MT : Makine Çevirisi (Machine Translation)
NER : Adlandırılmış Varlık Tanıma (Named Entity Recognition)
NLP : Doğal Dil İşleme (Natural Language Processing)
OCR : Optik Karakter Tanıma (Optic Character Recognition)
OP+SC : Spearman Korelasyonuyla Ortogonal Procrustes
OP : Ortogonal Procrustes (Orthogonal Procrustes)
PDF : Taşınabilir Doküman Formatı (Portable Document Format)
PPMI : Pozitif Noktasal Karşılıklı Bilgi
QA : Soru Cevaplama (Question Answering)
SC : Spearman Korelasyonu (Spearman Correlation)
SVD : Tekil Değer Ayrışımı (Singular Value Decomposition)
TBMM : Türkiye Büyük Millet Meclisi
TDK : Türk Dil Kurumu
TL : Türk Lirası
WER : Kelime Hata Oranı (Word Error Rate)

SEMBOL LİSTESİ

Bu çalışmada kullanılmış olan simgeler açıklamaları ile aşağıda sunulmuştur.

Simgeler	Açıklama
W	Kelime vektör matrisi
U	Sol tekil vektörler
Σ	Tekil değerler
$\mathbf{R}$	Dönüşüm matrisi
$\mathbf{Q}$	Ortonormal dönüşüm matrisi
$\mathbf{I}$	Birim matris
$\mathbf{w}$	Kelime vektörü
$\mathbf{E}$	İndeks olarak hizalanmış kelime vektör matrisi
$\mathbf{M}$	Doğrusal dönüşüm matrisi /
$\boldsymbol{u}$	Kelime / Düğüm
$\boldsymbol{u_i}$	Rastgele etki
$\boldsymbol{p}$	Olasılık Dağılımı
$\boldsymbol{\alpha}$	Katsayı (üstel ve doğrusal)
$\boldsymbol{u}$	Kelime rastgele etki
$\boldsymbol{p}$	Olasılık dağılımı
T	Zaman dilimi
c	Bağlam kelime
K	K en yakın komşuluk kümesi
TP	Doğru pozitif sayısı
TN	Doğru negatif sayısı
FN	Yanlış negatif sayısı
V	Kelime kümesi / Bir çizgedeki düğüm kümesi
E	Bağlantı Kümesi
$\boldsymbol{w}$	Kelime
$\boldsymbol{v}$	Kelime / Düğüm
V	Kelime kümesi / Bir çizgedeki düğüm kümesi
S	Eş anlamlısı olan bir kelime
x,y	Düğüm
C_B	Ara Merkezilik değeri

1. GİRİŞ

Diller, insan iletişiminin doğasını yansıtır. Çeşitli içsel ve dışsal faktörlerin etkisiyle zamanla evrilir. Teknolojik ilerlemelerle birlikte ortaya çıkan yeni konseptleri tanımlamak için yeni terimler dilde yer bulmaya başlar. Yalnızca teknolojik gelişmeler değil, kültürel değişimler farklı diller arasında kelimelerin benimsenmesine veya uyarlanmasına neden olur. Türkiye'de ve Türkçe'de belirgin bir şekilde görüldüğü üzere, siyasi kararlar da özellikle hükümetlerin veya kurumların dil kurallarını düzenleyen politikalar da dil değişimlerinde önemli bir rol oynar. Bu değişim, tamamen yeni kelimelerin yaratılması, modası geçmiş terimlerin yavaş yavaş ortadan kalkması ve kelimelerin orijinal anlamını kaybetmesi ya da daha genel bir kullanıma sahip olması gibi birçok farklı şekilde kendini gösterir. Bunların dışında dilin diğer katmanlarında da değişiklikler yaşanabilir; telaffuz kalıplarını değiştiren fonetik kaymalar, kelime oluşum kurallarını etkileyen morfolojik düzenlemeler başlıca değişen elementlerdir (Aitchison, 2001; McMahon, 1994; Bybee, 2015). Tüm bu faktörler, toplum ve dil arasındaki karmaşık ilişkiyi ortaya koyarak, dış ve iç faktörlerin dil değişimi içinde nasıl yer aldığını gözler önüne serer. Ancak, Türkçe'nin son yüz sene içindeki dönüşümü, diğer dillerde görülen doğal değişimlerden farklı olarak, devlet eliyle yönlendirilmiş benzersiz bir değişim sürecine sahne olmuştur. 1923 yılında Türkiye Cumhuriyeti'nin kurulmasının ardından, modernleşme yeni hükümetin temel hedeflerinden biri haline gelmiştir. Bu bağlamda, iki büyük dil reformu gerçekleştirilmiştir. İlk olarak Arap-Fars alfabesi bırakılarak 29 harfli Latin alfabesi benimsendi. Sonrasında dilin sadeleştirilmesi ve arındırılması amacıyla, Osmanlı Türkçesi'nde yoğun olarak kullanılan Farsça ve Arapça kökenli kelimelerin yerine, Türkçe kökenli kelimeler kullanılmaya başlanmıştır. Bu reformlar kapsamında, Türkçe'de sıklıkla neolojizm örnekleri görülmeye başlandı ve Türkçe söz varlığı köklü bir şekilde değişime uğradı. Türkçe'nin bu sistematik ve kasıtlı dönüşümü, dilin siyasi ve kültürel gündemler tarafından nasıl şekillendirilebileceğine dair dikkat çekici bir örnek olarak öne çıkmaktadır.

Türkçeye sistematik bir şekilde yapılan düzenlemeler dilin yapısında köklü değişimlere sebep olmuştur. Türkçe'nin ne denli değiştiğini anlamak için Mustafa Kemal Atatürk'ün Nutuk eseri incelenebilir. Nutuk'un 1927 yılında yazılmış versiyonu genç nesiller tarafından gitgide daha az anlaşılmaya başlanınca 1960'larda tekrar modern Türkçe olarak tekrar basılmıştır (Lewis G. 1999). Bu tez çalışmasında Türkçe'nin bu hızlı değişimini motivasyon alarak Türkçe'deki eş anlamlı kelimelerin dildeki değişimlerini diyakronik bir yaklaşımla, doğal dil işleme teknikleri kullanılarak incelenmiştir.

İlk olarak Türkçe diyakronik kaynak derlemini *Turkronicles*'ı oluşturduk. *Turkronicles,* Resmî Gazete ve Türkiye Büyük Millet Meclisi belgelerinden oluşturulmuş, 1920'den 2022'ye kadar olan dönemi kapsayan bir Türkçe diyakronik korpustur. *Turkronicles,* 45 bin belge ve 842 milyon token[1] içererek Türkçe'nin dilsel evrimini analiz etmek ve tarihi Türkçe belgeleri işlemek için modeller geliştirmek adına önemli bir kaynak. Buna ek olarak, diyakronik derlemler üzerinde hesaplamalı diyakronik analiz yapılabilmesini mümkün kılan bir kütüphane geliştirdik. Ayrıca, *Turkronicles'*ı kullanarak Türkçe kelime dağarcığının ve yazım kurallarının 1920'den bu yana nasıl değiştiği. Analizimiz, kelime dağarcığının önemli ölçüde değiştiğini ve bazı kelimeler için birden fazla yazım biçiminin bulunduğunu ortaya koymaktadır. Özellikle, beklendiği gibi, kelime dağarcığının farklılaşmasının zamanla arttığını gözlemlendi. Kelime dağarcığındaki bu önemli değişimlere rağmen, eski Türkçe kelimelerin yeni türetilmiş kelimelerle aynı anlamlara sahip olduğunu kelime vektör modelleri kullanarak tespit etmenin mümkün olduğunu gösterilmiştir. Yazım kurallarıyla ilgili olarak, şapka işaretlerinin kullanımında belirgin bir azalma tespit edildi. Spesifik olarak *b* ve *d* harfleriyle biten kelimelerin büyük ölçüde, sırasıyla *p* ve *t* harfleriyle biten karşılıklarıyla değiştirildiğini, ancak ilk biçimlerin halâ kullanımda olduğu gözlemlendi.

Diyakronik korpusu oluşturduktan sonra bu korpusu kullanarak farklı zaman dilimlerinde kullanılan eş anlamlı kelimelerin tespit edilmesi için iki farklı yöntem önerildi. İlk yöntemde, ilgili zaman dilimlerindeki belgelerden oluşturulan vektör uzaylarını hizalamak için Ortogonal Procrustes Problemi'nin çözümünü kullandık. İkinci yöntemde ise ilk yöntemi genişleterek, yıllar boyunca kelimelerin frekansları

[1] DDİ'de bir model tarafından işlenen bir metin birimi.

arasındaki Spearman Sıralama Korelasyonu dahil edildi. Ayrıca, hedef zaman dilimi 1960'lardan 1980'lere kaydığında önerilen yöntemlerin etkinliğinin tutarlı kaldığı gözlemlendi.

Son olarak eş anlamlı kelimelerin anlamsal değişmesinin anlam değişim mekanikleri ile nasıl uyum sağladığı araştırıldı. Spesifik olarak kelimelerin başka bir anlam alanına girdiği durumlarda, bu anlam alanına karşılık gelen diğer bir kelimede anlam daralması olup olmadığını eş anlamlı kelimeler üzerinden incelendi. Problem ilk olarak kelime ilişki çizgesi üzerinden analiz edildi ve anlam değişimlerini ölçmek için bir çizge merkeziyet ölçütü olan Ara Merkezilik değeri kullanıldı. Deneylerde incelenen eşanlamlı kelimelerin zaman serilerinin büyük çoğunluğunda hipotezi destekler nitelikte korelasyon katsayısı tespit edildi ve bunların çoğunun istatistiksel olarak anlamlı çıktığını gözlemlendi. Bu bulguları desteklemek amacıyla zaman serileri üzerinde bir Doğrusal Karma Model eğitildi ve bir kelimenin diğerine sabit etki katsayısının negatif işaretli olduğu bulundu. Bu sonuçlar, eşanlamlı kelimelerin anlamlarının birbirine zıt yönde değiştiğine dair kanıtlar sunmaktadır.

Bu çalışmada, Türkçe'nin tarihsel süreçteki değişimlerini analiz ederek dildeki değişimin etkilerinin gözlemlenmesi hedeflendi. Özellikle, eş anlamlı kelimelerin farklı dönemlerdeki kullanımları ve anlam kaymalarını incelenerek yapılan çalışmalar üç ana başlık altında ele alındı. Elde edilen bulgular, Türkçe'nin 20. yüzyıldan itibaren hızlı bir dönüşüm geçirdiğini ve dilin, tarihsel, politik ve kültürel etkilerle nasıl şekillendiğini ortaya koymaktadır. Bu tez çalışması, hem Türkçe'nin değişimi hakkında önemli veriler sunmakta hem de doğal dil işleme yöntemleriyle tarihsel metinlerin analizi için çeşitli yaklaşımlar önermektedir.

1.1 Literatür Taraması

Bu bölümde literatürde var olan çalışmalar üç farklı başlıkta ve bu tezin literatüre katkılarına paralel olacak şekilde incelendi.

1.1.1 Türkçe korpuslar ve diyakronik çalışmalar

Türkçe için Doğal Dil İşleme (NLP) kaynakları, İngilizce'ye kıyasla oldukça sınırlı olsa da, akademik çalışmalarda sıklıkla kullanılan birkaç Türkçe korpus bulunmaktadır. METU Korpus (Say ve diğ., 2002) ve Türkçe Ulusal Korpus (TNC) (Aksan ve diğ., 2012), farklı türlerden belgeler içeren genel amaçlı, tür dengeli

Türkçe korpuslardır. METU Korpus 2 milyon token içerirken, TNC, sözcük sınıfı etiketlemeleriyle 50 milyon kelimeye sahiptir. Her iki korpus da 1990'dan önce yazılmış belgeler içermemektedir. Ancak, bu veri setleri zaman kapsamları diyakronik analiz için bizim korpusumuzla karşılaştırıldığında oldukça kısıtlıdır. Literatür araştırmamıza göre Türkçe için erken cumhuriyet döneminden günümüze uzanan, nispeten büyük sayılabilecek, yalnızca bir diyakronik korpus bulunmaktadır. Bu korpus, 1920-2015 yılları arasındaki parlamento oturumlarına ait tutanaklardan oluşan TBMM Korpusudur (Güngör ve diğ., 2018). Bu dokümanlar milletvekillerinin konuşmalarının tam transkript versiyonları olduğundan, bu korpus modern Türkçe'nin tarihsel evrimini yansıtabilir. Bu tez çalışmasında ise, parlamento kayıtlarının zaman aralığını 2022'ye kadar uzatarak ve 1921-2022 yılları arasında yayımlanan Resmî Gazete sayıları taranarak daha kapsamlı bir korpus oluşturuldu. Böylece, literatür araştırmasına göre, *Turkronicles*, Türkçe'nin dil dinamiklerine ve Türkiye'nin siyasi söylemine dair tarihî bir bakış sunan en büyük diyakronik korpus olma özelliğini taşır.

Literatürde, dil değişiminin çeşitli yönlerini diyakronik bir perspektifle ele alan birçok çalışma bulunmaktadır. Michel ve diğ. (2011) ile Lieberman ve diğ. (2007) İngilizce'nin evrimsel dinamiklerine odaklanmışlardır. Esas olarak kelimelerin farklı zaman dilimlerindeki frekans analizi üzerine inşa edilen bu çalışmalar, tarih boyunca dil ile ilgili değişimleri incelemekte ve dilin değişimdeki uzun vadeli kalıpları ve kültürel değişimlerin etkilerini ortaya koymaya çalışmaktadırlar. Farklı bir metodolojik yaklaşım olarak, Pechenick ve diğ. (2015) İngilizce'nin evrimini incelemek için JSD gibi bilgi teorisi yöntemlerini sunulan Google Books veri seti (Michel ve diğ., 2011) üzerinde kullanmışlardır.

İngilizce için birçok diyakronik analiz çalışması bulunmasına rağmen, Türkçe için yapılan çalışmaların sayısı sınırlıdır. Bu çalışmalar, dilin çeşitli yönlerini diyakronik olarak incelemektedir. Örneğin, Salan ve Kabayı (2022) ile Vahit (2003), Türkçe kelimelerdeki kelime başı ünlülerindeki ses değişimlerine dair kapsamlı incelemeler sunmakta ve farklı dillerin sözlüklerini inceleyerek bu fenomenin örneklerini bulmaktadırlar. Sultanzade (2012), Dede Korkut Kitabı'ndaki bir dizi fiilin valans değişimini niteliksel ve niceliksel olarak incelemiş ve bunları modern Türkçe'deki karşılıklarıyla karşılaştırmıştır. Aksan (1965) ve Bahattin (2003) semantik değişimi incelemiştir. Hesaplamalı yöntemler kullanarak Türkçe'yi bir korpus üzerinden

diyakronik inceleyen araştırmalar da mevcuttur. Örneğin, Can ve Patton (2010), kelime uzunlukları, sıklıkları ve yazarların yazım stillerindeki değişikliklerini 1900'lerin başından 1990'ların sonuna uzanan aralıkta kapsayan edebî eserlerin oluşturduğu bir korpus üzerinde incelemiş ve istatistiksel modeller kullanarak ölçümlemiştir. Ayrıca, Altıntaş ve diğ., (2007) tarafından yapılan çalışma, edebi eserlerin farklı zaman aralıklarındaki çevirilerini karşılaştırarak kelime köklerinin uzunluğunun zamanla azaldığını, eklerin uzunluğunun ise zamanla arttığı sonucuna ulaşmıştır. Araştırmalarımıza dayanarak, Güngör ve diğ. (2018) tarafından yapılan çalışma, Türkçe dilini büyük korpus tabanlı diyakronik bir şekilde analiz eden bu çalışma, bu tez çalışmasına en yakın çalışmadır. Bu çalışma, yakın eşanlamlı kelimelerin sıklıklarındaki değişimleri ve konu dağılımlarını LDA yöntemini kullanarak Türkçe'nin dil değişimini gözler önüne serer. Bu tez çalışmasında ise, daha büyük bir korpus üzerinde kelime dağarcığı ve yazım kurallarındaki değişimi gözlemlemek için farklı hesaplamalı yaklaşımlar kullanılmıştır.

1.1.2 Eş anlamlı kelimelerin diyakronik tespiti

Neolojizm, yeni kelimeler veya kavramlar yaratmayı ifade eder ve genellikle dil içi, dil dışı ve semantik mekanizmalar tarafından yönlendirilir (Rodríguez Guerra, 2016). Dilbilimde, neolojizm önemli bir kavramdır çünkü dillerin kültürel ve teknolojik değişimlere nasıl uyum sağladığını ortaya koyan mekanizmaları aydınlatır (Taylor, 2015). Ayrıca, neolojizm anlayışı, dilbilimcilere diller arasındaki etkileşim bağlamında dil değişimini ve evrimini izleme imkânı tanır (Bybee, 2015). Literatürde, neolojizme hesaplamalı bir perspektiften yaklaşan çalışmalar bulunmaktadır. Pechenick ve diğ. (2015) ve Michel ve diğ. (2011) İngilizce üzerinde diyakronik bir analiz gerçekleştirerek, İngilizce'yi on yıllık dilimlerle karşılaştırarak dildeki değişiklikleri frekans tabanlı yöntemlerle gözlemlemişlerdir.

Dağılımsal anlam yöntemlerinin uygulanmasıyla, vektör temsilleri (Mikolov ve diğ., 2013), DDİ çalışmalarında etkili bir şekilde kullanılmış ve çeşitli dilbilimsel hipotezlerin test edilmesine olanak tanımıştır (Xu ve Kemp, 2015; Ryskina ve diğ., 2020). Kelime vektörlerinin sıklıkla kullanıldığı alanlardan biri diyakronik analizdir. Semantik değişimde, genellikle bir diyakronik korpus belirli zaman dilimlerine ayrılır ve her bir dönem için farklı türdeki kelime gömme algoritmaları kullanılarak bir vektör alanı elde edilir (Kulkarni ve diğ., 2015; Hamilton ve diğ., 2016; Rudolph

ve Blei, 2018). Kelime vektörleri oluşturulduktan sonra, bu alanlar farklı zaman dilimlerinde vektör karşılaştırmaları yapmak için hizalanır. Vektör alanlarını hizalamak için çeşitli yöntemler vardır, bunlar arasında doğrusal dönüşümler uygulama (Kulkarni ve diğ., 2015), Ortogonal Procrustes hizalamasıyla ortogonal dönüşüm matrisi bulma (Hamilton ve diğ., 2016) ve sorgu kelimesinin vektörünü, hedef zaman dilimindeki ilişkili kelimeleri dahil ederek hizalama (Zhang ve diğ., 2015) yer almaktadır. Başka bir yaklaşım olarak, bazı araştırmacılar, hizalamayı eğitim aşamasında dikkate alarak kelime vektörlerini ortak bir şekilde eğitmeyi düşünmüşlerdir (Yao ve diğ., 2018; Di Carlo ve diğ., 2019). Bu tez çalışmasında, Ortogonal Procrustes Hizalaması kullanılmıştır. Ayrıca, Spearman Korelasyon Katsayısı'nı temel alan bir sıralama mantığını entegre ederek önerilen yöntem başarısı arttırılmıştır.

Birçok araştırmacı, kelime vektörlerini kullanarak neolojizmin etkilerini ele almıştır. Zhang ve diğ. (2015) yerel ve global karşılık eşleştirmesi yaparak bir sorgu kelimesinin karşılığını bulmuşlardır. Global karşılık yaklaşımında, tohum kelimelerinin zamansal temsilleri arasındaki Öklid mesafesini minimize ederek bir dönüşüm matrisi elde etmişlerdir. Yerel karşılık yaklaşımında ise, sorgu kelimesinin karşılığını bulmak için semantik bir çizge temsili oluşturmuşlardır. Ancak, Zhang ve diğ. (2017), aynı global dönüşümün tüm kelimelere uygulanması durumunda performans kaybı gözlemlemişlerdir. Her bir kelime için dönüşüm matrisi oluşturarak global dönüşüm yaklaşımının performans kaybını aşmışlardır. Bu çalışmalar, analojik karşılıklar bulmayı amaçlamaktadır, örneğin, 1987-1991 yıllarından *Walkman*[2] kelimesini bulmak için 2002-2007 yıllarından iPod kelimesinin kullanılması (Zhang ve diğ., 2015). Bu, bizim durumumuzdan farklıdır çünkü analojik karşılıklar bulmak, eşleşen kelimelerin benzer veya ilişkili semantik alanlara ait olmasını gerektirir; zira sorgu kelimelerinin tam anlamı önceki zaman dilimlerinde ortaya çıkmamış olabilir. Buna karşılık, bu tez çalışmasında, Türk Dil Devrimi nedeniyle modern Türkçe kelimelerle değiştirilmiş, genellikle Arapça ve Farsçadan alınan kelimelere odaklanılmıştır. Bu nedenle amaç, belge-vesika gibi tam olarak aynı kavram veya anlamla karşılık gelen kelimeleri bulmaktır. Bir diğer fark ise, tüm bu çalışmaların İngilizce üzerinden yürütülmesidir. Bildiğimiz kadarıyla,

[2] Müzik çalar.

çalışmamız, Türkçe'de neolojizmi hesaplamalı bir perspektiften inceleyen ilk çalışmadır.

1.1.3 Anlam değişimi

Literatürde semantik değişim uzun yıllardan beri tarihsel dilbilimciler tarafından incelenen bir alandır. Bréal (1897) semantik kavramını tanımlarken aynı zamanda insan dilini şekillendiren kuralları semantik açıdan inceledi. Ardından değişimleri kategorilerine ayırıp değişim sebepleri için öngörüler ortaya koydu. Semantik kavramının ile ilgili öğretilerin ortaya çıkmasından sonra yapısal yaklaşımların ile semantik değişim üzerindeki rolü araştırmacıların dikkatini çekti ve anlamsal değişimleri araştıran çalışmalar anlam alanları üzerinden incelenmeye başlandı (Stern, 1931; Traugott, 1985). Son çalışmalarda odaklanılan nokta ise anlam değişiminin dil bilimsel sebepleri ve mekanizmalarının yanında bu olgunun bilişsel süreçlerine dayanır (Traugott, 1985). Bu bilişsel süreçlere örnek olarak konuşmacının ve dinleyicinin konuşmadaki rolü, analojik düşünce gibi faktörler sayılabilir.

Vektör gösterimlerinin ortaya çıkması ve büyük ölçekli korpusların erişiminin kolaylaşmasıyla anlam değişimi doğal dil işleme komünitesinin odağına girmiştir. DDİ alanında yürütülen semantik değişim çalışmalarında, diyakronik bir korpus genellikle belirli zaman dilimlerine ayrılarak her dönem için farklı kelime vektör algoritmaları ile bir vektör alanı oluşturulur (Kulkarni ve diğ., 2015; Hamilton ve diğ., 2016; Rudolph ve Blei, 2018). Bu vektör alanları, kelimelerin farklı dönemlerdeki değişimlerini analiz etmek amacıyla hizalanarak karşılaştırılır.

Anlam genişlemesi ve daralması konusunda literatürde vektörleri ve çizgeleri kullanan çalışmalar mevcuttur. Bu konuda bazı çalışmalar ortak-oluşum matrislerine bağlı oluşturulan vektörlere dayanır (Sagi ve diğ., 2009; Tang ve diğ., 2016). Daha sonra bu kelimelerin zaman boyunca ilgili ölçütlerle vektörel değişimlerine bakarak anlamın ne kadar daraldığı veya genişlediği ölçülmeye çalışılır. Burada vektörlerin bileşenleri oluşturulurken farklılıklar görülebilir. Örneğin, Sagi ve diğ. (2009) doğrudan ortak oluşum vektörlerini kullanırken Tang ve diğ. (2016) ise vektörlerde entropideki değişime bakarak genişlemeye veya daralmaya karar vermiştir. Farklı bir yaklaşımda ise doğrudan ortak-oluşum matrisi bir çizge olarak görülüp bu çizge üzerinde kümeler birer anlam alanı olarak düşünülmektedir. Ardından bu kümelerin değişimi zaman içinde takip edilip anlam alanlarının ne tür bir anlamsal değişime

uğradığı tespit edilir (Mitra, ve diğ., 2015). Cassotti ve diğ., (2024) bağlamsal kelime vektör modelleriyle (RoBeRTa-Large), WordNet tabanlı senkronik anlam ilişkilerinden yararlanarak kelimeler arasındaki anlamsal ilişkileri sınıflandırmış ve bunları genişleme ve daralma dahil olacak şekilde anlam değişim türleriyle eşleştirmiştir.

Önceki çalışmalar incelendiğinde İngilizce için yapılan pek çok araştırma olduğu görülüyor. Literatür araştırmamıza göre Türkçe'de henüz hesaplamalı olarak anlam genişlemesi ve daralmasını inceleyen bir çalışma bulamadık. Bu çalışmada diğerlerinden farklı olarak Türkçe dilinde anlam genişlemesi ve daralmasını incelendi. Ayrıca yöntem ve ölçüt olarak diğer çalışmalardan farklı bir şekilde anlamsal değişim modellendi.

2. TURKRONICLES: TÜRKÇE İÇİN DİYAKRONİK KAYNAKLAR

Yaşayan her canlı varlık gibi, diller de zamanla pek çok farklı perspektiften değişime uğrar. Bu değişimlerin temelinde teknolojik gelişmeler, sosyal değişimler, kültürel geçişler ve siyasi kararlar gibi birçok etken vardır. Dilin kelime hazinesine yeni kelimelerin eklenmesi, kelimelerin zamanla yok olması, anlam, ses yapısı, biçim ve dilbilgisi kurallarında yaşanan değişimler dillerin geçirdiği değişimin başlıca kategorilerini oluşturur (McMahon, 1994; Aitchison, 2001; Bybee, 2015).

Türkçe, son yüzyılda diğer dillere kıyasla dikkate değer bir şekilde farklı bir değişim yolunu takip etti. 1923'te Türkiye Cumhuriyeti'nin kurulmasından sonra, kültürel ve teknolojik modernleşme, yeni hükümet tarafından benimsenen reform programının en önemli gündemlerinden biri olmuştur. Türk dili, bu modernleşme süreci çerçevesinde iki köklü değişim geçirmiştir. Bu değişimler ilk olarak dilin temel bileşenleri olan alfabe sistemini ve sözcük dağarcığı üzerinde etkilerini göstermiştir. 1928'de, Osmanlı İmparatorluğu döneminde kullanılan Arap-Fars alfabesi kaldırılarak, 29 harften oluşan Latin alfabesine geçildi. Türkçe'deki ikinci önemli değişim ise Osmanlı döneminde çok sayıda bulunan yabancı kökenli kelimelerin tarihsel olarak Türkçe olan ya da Türkçe'nin morfofonolojik[3] kuralları ile türetilebilen kelimelerle değiştirilmesi yoluyla Türk dilini "sadeleştirme" ve "saflaştırma" girişimiydi.

Türk dil reformu sürecinde yeni türetilmiş kelimeler veya ifadeler, yani neolojizmlerin (Fischer, 1998) yoğun bir şekilde dilde görülmesiyle, kelime dağarcığında önemli değişikliklere yol açmıştır. Bu durum, özellikle genç nesiller için dil reformu öncesine ve ilk yıllarına ait metinleri anlamak güçleştirmiştir Ayrıca, dilin evrimi, Doğal Dil İşleme (DDİ) modellerinin tarihsel metinlere uygulanmasında çeşitli zorluklar ortaya çıkarmaktadır. Bu metinler üzerinde işlem yapmak için kullanılan modeller ile zaman uyumsuzluğu (Luu ve diğ., 2022) gösterir. Çünkü DDİ modelleri eğitim verisinde bulunmayan tarihsel ifadeler ile karşılaştıklarında, Soru Yanıtlama (QA) (Zheng ve diğ., 2024) ve Adlandırılmış Varlık Tanıma (NER) (Luu ve diğ., 2022; Onoe ve diğ., 2022) gibi çeşitli görevlerde performansları düşme eğilimindedir. Örneğin, Zheng ve diğ. (2024), metinde neolojizmler bulunduğunda

[3] Morfemlerin fonolojik temsili.

Makine Çevirisi (MT) performansının düştüğünü bildirir. Bu nedenle, tarihî belgeleri etkili bir şekilde analiz edebilmek için veri kaynaklarına ve modellere ihtiyaç duyulmaktadır. Ayrıca Türkçe içerisinde dil değişimi inceleyen korpus tabanlı çalışmalar oldukça kısıtlıdır. Literatür taramasında Türkçe için büyük ölçekli korpus üzerinden dil değişimini inceleyen yalnızca bir diyakronik korpus bulunmaktadır (Güngör ve diğ., 2018).

Bu bölümde, Türkçe'de var olan en büyük diyakronik korpus *Turkronicles'ı* oluşturduk ve bu korpus aracılığıyla 1920'lerden bu yana Türkçe'nin nasıl değiştiğini incelendi. İlk olarak 1920-2022 yılları arasını kapsayan Resmî Gazetesi ve Türkiye Büyük Millet Meclisi (TBMM) tutanakları toplandı. Sonuç olarak, 45.375 belge, 842 milyon kelime ve 211 bin özgün kelimeden oluşan Turkronicles korpusu oluşturuldu. Ayrıca, çekimli durumda ve lemmatize[4] edilmiş kelimelerin için n-gram frekansları ölçüldü. Bunun yanı sıra, 1920'den bu yana Türkçe kelimelerin anlamsal değişimlerini araştırmaya olanak tanımak amacıyla her on yıllık dönem için kelime vektör modelleri oluşturuldu. Dahası, Türkçe'de bulunan modern ve tarihî kelime eşlerinden oluşan bir karşılıklar sözlüğünü otomatik olarak dijitalleştirip kullanıma açıldı. Son olarak, araştırmacıların diyakronik korpuslarla dilbilimsel analizler yapmalarına yardımcı olmak amacıyla Lingan adında bir Python kütüphanesi kullanıma sunuldu.

Korpusta bulunan her iki kaynağın da yasalar, düzenlemeler ve bunlarla ilgili tartışmalar gibi politik eylemler hakkında belgeler içerdiğini göz önünde bulundurarak, diyakronik korpusun Türk dilinin evrimi ve hükümetin bu dönüşümdeki rolünü analiz etmek için önemli bir kaynak olduğu düşünülebilir. Bu nedenle, korpusumuzu kullanarak aşağıdaki araştırma sorularına (AS) yanıt aranmaktadır.

AS-1: 1920'den bu yana Türkçe kelime dağarcığı nasıl değişti? Veri kümesi on yıllık dönemlere ayrıldı ve her bir zaman periyodu için kullanılan kelimeleri farklı metodolojilerle karşılaştırıldı. Zaman farkı arttıkça iki dönem arasındaki kelimelerin daha fazla farklılaştığı gözlemlendi. Beklendiği üzere, yeni türetilen Türkçe kelimelerin sıklığı zamanla artarken, yabancı kökenli kelimelerin sıklığının azaldığını görüldü. Geçen yüzyıl boyunca dildeki büyük değişikliklere rağmen,

[4] Sözlük formuna indirgeme.

oluşturulan diyakronik kelime vektör modelleri ile aynı veya benzer anlamlara sahip Arapça veya Farsça kökenli kelimeleri yeni türetilen Türkçe kelimelerle eşlemenin mümkün olduğu gösterildi.

AS-2: 1920'lerden bu yana yazım kuralları nasıl değişti? 1920'ler ve 1930'larla kıyaslandığında, şapka işaretinin kullanımının belirgin şekilde azaldığı gözlemlendi. Ayrıca, son harfi *b* olan kelimelerin (örneğin, *kitab*) kullanımı, son harfi *p* olan formuna (yani *kitap*) kıyasla zamanla azaldığı görüldü. Ancak, son harfi '-d' veya '-t' olan kelimeler için farklı bir örüntü fark edildi: 2010'larda son harfi '-d' olan kelimelerin son harfi '-t' olan kelimelere (örneğin, Ahmed vs. Ahmet) oranı, 1920'lerdeki oranına benzer olduğu gözlemlendi. Bununla birlikte, 1990'lardan bu yana bu oranın azalan bir eğilim gösterdi.

2.1 Resmî Gazete

Resmî Gazete, hükümet faaliyetleri ve devlet adamlarının çeşitli konular hakkındaki görüşleri gibi diğer konularda bilgi vermek amacıyla 7 Ekim 1920'de kurulmuştur. Yayın sıklığı düzenli aralıklardan (haftalık, günlük vs.) düzensiz yayınlara değişiklik gösterir. Günümüzde, tatil günleri ve pazar günleri hariç her gün yayımlanmaktadır.

Çizelge 2.1 : 1920 ve 2020 yılları arasındaki Resmî Gazete'den örnek cümleler.

Tarih	Cümleler
7 Şubat 1921	(1) Hâkimiyet bilâ-kayd ü şart milletindir. İdare usulü, halkın mukadderatını bizzat ve bil-fiil idare etmesi esasına müstenittir. (2) Türkiye Devleti, Büyük Millet Meclisi tarafından idare olunur ve hükûmeti "Türkiye Büyük Millet Meclisi Hükûmeti" unvanını taşır.
4 Ekim 2020	(1) Bu yönetmeliğin amacı, TOBB Ekonomi ve Teknoloji Üniversitesi Laboratuvar Okullarındaki eğitim-öğretim, yönetim, kayıt-kabul, devam-devamsızlık, nakil ile öğrenci başarısının tespiti ve işleyişe yönelik usul ve esasları düzenlemektir. (2) Laboratuvar okulları ile Üniversitenin öğrenci, öğretim elemanı ve öğretmenleri birbirlerinin dersleri ile kültür, sanat, spor ve sosyal faaliyetlerine katılarak müşterek etkinlikler gerçekleştirirler.

Resmî Gazete'nin içeriği Türkiye'nin yönetim sürecinin bir yansımasıdır. Resmî Gazete sayıları çoğunlukla aşağıdaki konularla ilgili haberleri içerir:

- Kanunlar,

- Türkiye Büyük Millet Meclisi kararları ve iç tüzükleri
- Uluslararası antlaşmalar,
- Cumhurbaşkanı yardımcısı, yüksek yargı üyeleri ve bakanlar gibi yetkililerin görevden alınması, seçimi, atanması, yerine başkasının atanması veya istifası ile ilgili işlemler,
- Cumhurbaşkanı yardımcısı, yüksek yargı üyeleri ve bakanlar gibi yetkililerin görevden alınması, seçimi, atanması, yerine başkasının atanması veya istifası ile ilgili işlemler,
- Cumhurbaşkanı tarafından yapılan atama, görevden alma ve görev sonlandırma kararları,
- İdari yargı değişiklikleri ve belediye kurulmasına ilişkin kararlar gibi iç idari kararlar.

Resmî Gazetede yer alan belgelerde neredeyse hiç yazım veya dilbilgisi hatası içermeyen resmi bir dil kullanılmaktadır. Uluslararası anlaşmaların yayımlanma zorunluluğundan dolayı belgelerde Türkçe olmayan metinler de bulunabilir. Ayrıca, ilk 1053 sayı, orijinal olarak Osmanlı alfabesiyle yazılmıştır. Ancak, 1928'de Harf Devrimi'yle Latin harflerini kullanılmaya başlanmıştır. Resmî Gazete dokümanları figürler, tablolar ve benzeri cümle dışı yapılar içerebilir. Çizelge 2'de, Resmî Gazete'de yer alan örnek cümleler verilmiştir.

2.2 Genişletilmiş TBMM Tutanakları

Güngör ve diğ. (2018) 1920-2015 yıllarını kapsayan korpusunun zaman aralığını 1920-2022 olarak genişlettik. TBMM tutanakları 1920'den itibaren sistematik olarak belgelenen Türkiye Büyük Millet Meclisi oturumlarında gerçekleşen tüm faaliyetlere dair metinlerden oluşmaktadır; bu faaliyetler arasında konuşmalar, denetimler, oylamalar, gürültüler, tartışmalar, programlar, gündemler, raporlar, mektuplar ve öneriler bulunmaktadır. Parlamentonun yıllık toplantı takvimindeki değişiklikler nedeniyle, bu belgelerin yayın sıklığı Resmî Gazete'ye kıyasla düzensizdir.

TBMM tutanakları ve Resmî Gazete, çizelgeler, tablolar ve ele alınan konular gibi yapısal unsurlar açısından önemli ölçüde benzerdir. Ancak, Resmî Gazete'nin aksine,

meclis kayıtları konuşmacıya ve bağlama bağlı olarak resmi dilden günlük konuşma diline kadar değişen bir yelpazeyi içerir. 1920-1928 yıllarına ait belgeler, Resmî Gazete'de olduğu gibi Osmanlı alfabesiyle yazılmıştır. Ancak, bu belgelerin harf çeviri versiyonları resmi internet sayfalarında bulunmaktadır.

Belgeleri toplamak için açık kaynak kodlu bir Python web tarama kütüphanesi olan *Scrapy* kullanıldı. Dokümanların kendisiyle beraber ilgili web sayfalarından metadataları[5] elde edildi. Metadata bilgisi yayımlayan kişi veya kurum, yayımlanma tarihi, dosya adı, indirme bağlantısı ve cilt bilgilerini içermektedir. Sonuç olarak, 7 Şubat 1921 ile 31 Aralık 2022 arasında yayımlanmış 31.999 adet Resmî Gazete sayısı ve TBMM tutanakları içinse 23 Nisan 1920 ile 1 Ağustos 2022 tarihleri arasında yayımlanmış 13.376 belge derleyip bir korpus haline getirildi.

2.3 Önişleme

Turkronicles korpusunda bulunan dokümanların çoğu *.pdf* formatında bulunmaktadır. İlk olarak indirilen dokümanları kolayca işleyebilmek için bu dokümanları *PyPDF* kütüphanesi aracılığıyla *.txt* formatına dönüştürdük. Fakat Resmî Gazete'nin 24.092 ile 28.500 sayılarının içerikleri *.pdf* dosyası yerine doğrudan *HTML* metin formatında olarak paylaşıldığı için bu aralıktaki dokümanları ilgili web sayfalarından doğrudan elde edildi.

Elde edilen metin dosyalarını manuel olarak inceledik. İncelememiz sonucunda dokümanlarda gürültü varlığı tespit edildi. Bu gürültü bileşenleri kaynağı çeşitli nedenlerle ortaya çıkmaktadır: Dokümanların fiziksel durumunun zarar görmesi, cümle dışı (tablo, figür vs.) bileşenlerin dokümanlardaki varlığı, Optik Karakter Tanıma (OCR) sistemlerinin sınırlamaları nedeniyle oluşan hatalardır. Bu sebeple aşağıdaki önişleme adımları gerçekleştirildi:

i) Birden fazla ardışık boşluğu tek bir boşluğa indirgendi. Ayrıca farklı türde boşluk karakterlerini standartlaştırıldı

[5] Veri hakkında veri.

ii) Tokenization[6] işleminde problemlere neden olan ve/veya UTF kodlamasında bir temsili olmayan karakterleri (örneğin \xa0 ve \xad) silindi.

iii) Gözlemlere dayanarak, dokümanlardaki gürültü ifadelerin nadir görülen tokenlerin[7] oluşmasına sebebiyet verdiğini gözlemledik. Bu nedenle sıklığı belirli bir eşik değerden daha az görülen tokenler elendi. Ancak, .pdf dosyalarının kalitesinin ve dolayısıyla *PyPDF*'nin performansının yıllar içinde değiştiği fark edildi. Bu nedenle, tüm belgeler için tek bir eşik değeri kullanmak yerine, her bir zaman dilimi için eşik değerini $\left\lceil \frac{N}{10000000} \right\rceil$ formülüyle belirledik. Burada N bir zaman periyodundaki toplam token sayısını ifade etmektedir. Bu tarz bir filtreleme yönteminin gürültülü kelimeleri filtrelediği gözlemlendi.

iv) Türkçe sondan eklemeli ve morfolojik olarak zengin bir dil olduğundan çekimli kelimelerin doğrudan analizinde yanıltıcı bir etkiye sahip olabilir. Bu sebeple kelimeleri sözlük formuna indirgemek için, Öztürel ve diğ. (2019) tarafından önerilen morfolojik analiz aracı kullanıldı. Bu aracın kelimeleri analiz etmekte yetersiz kaldığı noktalarda ise Can ve diğ. (2008) çalışmasında kök bulmada etkili olduğu gösterilen F5 (kelimenin ilk beş harfini kök olarak belirleme) yöntemi uygulandı.

2.4 Dijitalleştirilmiş Karşılıklar Sözlüğü

Eski Osmanlıca kelimelerin modern Türkçe karşılıklarının bulunduğu bir karşılıklar sözlüğü Türk Dil Kurumu (TDK) tarafından 1935 yılında yayımlandı (TDK, 1935). Bu sözlük, 1930'lu yıllarda modern Türkçe'ye yeni girmiş kelimelerin Osmanlıca karşılıklarını belirli bir düzende okuyucuya sunar. Sözlükteki her bir girdi, çok anlamlı kelimelerin farklı anlamlarını, eşanlamlıları, terimleri, kısaltmaları ayıran özel işaretlerle belirtilmiştir.

[6] Cümlelerin modellerin işleyebileceği nispeten anlamlı birimlere bölünmesi işlemi.
[7] Tokenization işlemi sonucunda ortaya çıkan birimlere verilen isim

Bu sözlüğü dijitalleştirmek adına ilk olarak *.pdf* dokümanını *tesseract* Optik Karakter Tanıma (OCR) kütüphanesi ile düz metin dosyasına dönüştürüldü. Ardından, yeni türetilmiş Türkçe kelimeleri ve eski karşılıklarını çıkarmak için kural tabanlı yöntemler (örneğin Düzenli İfadeler) kullanıldı. Nihayetinde 8.647 girdi içeren bir JSON dosyası oluşturuldu.

2.5 n-gram'lar

n-gram'lar[8], araştırmacılar tarafından dil değişimini ve kültürel değişimi analiz etmek için çeşitli çalışmalarda kullanılmıştır (Pechenick ve diğ., 2015; Michel ve diğ., 2011). Bu bağlamda n-gram'ların tarih boyunca istatistiklerinin sunulduğu en büyük ve kapsamlı sayılabilecek araç Google Books Ngram Viewer'dır[9]. Fakat bu platformda Türkçe dili şu an kullanılabilir değil. Bu nedenle, korpusumuzdan elde edilen n-gram ile alakalı dosyaları diğer tüm kaynaklarla birlikte erişilebilir hale getirildi. Bilhassa, her dosyadan frekanslarıyla birlikte 1-gram, 2-gram ve 3-gram verisi kaydedildi. Ek halindeki kelimeler ve sözlük formuna indirgenmiş kelimelerin frekansı hesaplandı. Ayrıca n-gram'ları zaman dilimlerine göre gruplandı. Böylece belirli kelime veya ifadelerin kullanımındaki değişimlerin incelenmesi mümkün kılındı. Bunun yanı sıra, kelimeler arasındaki kullanım ilişkisini daha derinlemesine anlamak için unigramların[10] PPMI değerlerini de verilere eklendi. Bu kaynaklar, araştırmacıların zaman içinde ifadelerin ve kelimelerin frekans değişimlerini kolayca incelemelerine, dilbilimsel kalıpları keşfetmelerine ve farklı zaman dilimlerinde Türk dili hakkında hipotezler test etmelerine olanak tanıyacaktır.

2.6 Diyakronik Kelime Vektörleri

Diyakronik kelime vektörleri DDİ çalışmalarında kelimelerin anlamlarının zaman boyunca nasıl değiştiğini, daha teknik bir tabirle semantik kayma, incelemek için kullanılan bir yöntemdir (Hamilton ve diğ., 2016; Kulkarni ve diğ., 2015). Bu nedenle, oluşturulan diyakronik kaynak derlemi içine Turkronicles belgeleri üzerinde eğitilmiş diyakronik kelime vektörleri dahil edildi. Bu vektörler üç farklı kelime

[8] Metin verisinde yan yana geçen N ardışık birim.
[9] https://books.google.com/ngrams/
[10] 1-gram

vektör algoritması ile oluşturuldu: Pozitif Noktasal Karşılıklı Bilgi (PPMI) yöntemi, PPMI matrisinin Tekil Değer Ayrışımı (SVD) ve Sürekli Kelimeler Torbası (CBOW) (Mikolov ve diğ., 2013). Öncelikle, ön işlenmiş metin dosyaları yayımlanma tarihlerine göre 10 yıllık dönemlere ayrıldı. Her dönem için terimden terime eşleşmeleri sayarak PPMI matrisini oluşturuldu. Bu matrisin elemanlarını doldurmak için Eşitlik (2.1)'de gösterilen formülden yararlanıldı:

$$PPMI(u,v) = max\left(log\frac{p(u,v)}{p(u)p_\alpha(v)}, 0\right) \quad (2.1)$$

Burada u ve v birer kelime, $p(u,v)$ bu kelimelerin ortak olasılık dağılımı ve $p(u)$ u'nun marjinal olasılık dağılımına karşılık gelmektedir. PPMI değerlerinin nadir olaylara hassas olduğu bilinmektedir (Mikolov ve diğ., 2013). Bu durumdan kaynaklanan negatif etkileri azaltmak adına $p_\alpha(v)$ v'nin düzeltilmiş unigram dağılımı PPMI değerlerinin hesaplanmasına dahil edilir. Deney aşamasında α değerini 0,75 olarak belirlendi. Ayrıca hedef kelimeleri bağlam kelimeleriyle ilişkilendirmek için boyutu 5 olan bir bağlam penceresi kullanıldı.

SVD yaklaşımında, kelimelerin vektör gösterimlerini $\mathbf{W} = \mathbf{U\Sigma^{1/2}}$ ile hesapladık; burada $\mathbf{U}$ matrisi PPMI matrisinin sol tekil vektörleri ve $\mathbf{\Sigma}$ ise PPMI matrisinin tekil değerleridir. Vektör boyutu olarak 300 değeri seçildi. Klasik SVD uygulamasından farklı olarak kelime vektörleri tekil değerler matrisinin karekökü alınarak hesaplanır. Bu işlemin kelime vektörlerinin kalitesini arttırdığı gösterilmiştir (Levy ve diğ, 2015).

Son olarak, her bir zaman dilimi için CBOW vektörlerini *gensim* kütüphanesi kullanarak oluşturduk. CBOW algoritması için SVD gömmelerinde olduğu gibi 2 boyutunda bir bağlam penceresi, α değeri 0,75 ve vektör boyutunu 300 olarak seçtik. Ayrıca, negatif kelime sayısı ve düşürme oranı gibi CBOW'a özgü parametreleri sırasıyla 5 ve 0,00001 olarak belirledik.

SVD'nin tekil olmayan yapısı ve CBOW algoritmasında yer alan rastgele süreçler, farklı zaman dilimlerine ait olan vektörlerin karşılaştırılmasını engeller (Hamilton ve diğ., 2016; Kulkarni ve diğ., 2015). Ancak, iki vektör uzayı bir dönüşüm matrisi $\mathbf{R}$ aracılığıyla döndürülebilir, ötelenebilir veya ölçeklenebilir. Bu nedenle, farklı zaman dilimlerine ait kelime vektörlerini karşılaştırmak için kelime vektör matrislerinin

$\mathbf{W_{t_1}}$ve $\mathbf{W_{t_2}}$'yi birbirine hizalamak gerekir. Bu sebeple hizalama algoritması olarak Ortogonal Procrustes yöntemi kullanıldı (Schönemann, 1966; Hamilton ve diğ., 2016). Ortogonal Procrustes probleminin çözümü, iki vektör uzayının kare farkının Frobenius uzaklığını en aza indiren bir dönüşüm matrisi $\mathbf{R}$ bulmayı amaçlar. Bu ifadenin probleme uygulanmış halinin matematiksel gösterimi Eşitlik (2.2)'de verilmiştir.

$$\boldsymbol{argmin}||\mathbf{W_{t_1}R} - \mathbf{W_{t_2}}||_{\boldsymbol{F}} \tag{2.2}$$

Ayrıca $\mathbf{R}$ üzerinde bir ortogonal olma kısıtı bulunmaktadır. Yani transformasyon matrisi $\mathbf{R^T R = I}$ koşulunu sağlamalıdır. $\mathbf{R}$ matrisi bulunurken ilk olarak $\mathbf{M = W_{t_2}}^{\mathbf{T}}\mathbf{W_{t_1}}$matrisinin SVD kullanılarak sol tekil vektörleri $\mathbf{U}$ ve sağ tekil vektörleri elde edilir. Ardından $\mathbf{R = UV^T}$ işlemiyle transformasyon matrisi oluşturulur. Bu yöntem, farklı zaman dilimlerindeki kelime vektörlerinin karşılaştırılmasını mümkün kılar. Bu yöntemi kullanarak farklı zaman dilimlerinin hizalanmış vektörleri ve her zaman dilimi çifti için bir dönüşüm matrisi kaynaklara dahil edildi.

2.7 Lingan: Diyakronik Analiz için Python Kütüphanesi

Diyakronik korpuslar üzerinde dilbilimsel analizler yapmak amacıyla *Lingan* adlı genel bir Python kütüphanesi geliştirildi. Bu kütüphaneyi geliştirerek diyakronik dilbilimsel analize hesaplamalı yaklaşım yöntemlerini kullanan araştırmaların deney aşamasını daha kolay uygulanabilir hale getirmek amaçlandı. Literatür taramasında diyakronik analiz için geliştirilen kütüphanelerin oldukça az olduğu gözlemlendi. Daha spesifik olarak literatürde yalnızca bir adet açık kaynak Python kütüphanesine rastlandı, *DRIFT* (Sharma ve diğ., 2021). *Lingan*'ın, *DRIFT* kütüphanesinden farkı ise hesaplamalı diyakronik analiz açısında farklı fonksiyon ve operasyonlara sahip olmasıdır.

Lingan, temel olarak üç farklı soyutlama katmanına dayanır: *Data, Container* ve *Operation.* Çizelge Ek.1, bu bileşenler üzerinde tanımlı fonksiyon ve metotları ayrıntılarıyla açıklamaktadır.

Data bileşeni, dilbilimsel analizde kullanılan gerçek verilerin bir temsilidir. Bu bileşenin temel sorumluluğu, veriyi yönetmektir. Pratikteki karşılıkları, kelime vektörleri, kelime haznesi ve n-gramlar gibi çeşitli veri türlerini içerir. Bu veri

türlerinin programlama dilindeki karşılıkları olan ve *Data* sınıfından türeyen *Embeddings*, *NGrams* ve *Vocabulary* sınıfları da kütüphanenin içine dahil edildi. Bu sınıflar, gerçek verilerin çeşitli DDİ görevlerinde etkili bir şekilde temsil edilerek işlenmesini kolaylaştırır. *Container* topluca oluşturulan metinsel verileri temsil eder ve teknik olarak, verileri temsil eden nesneleri sarmalamak için kullanılan konteyner sınıflarının hiyerarşik yapısı için ortak bir arayüzdür. *Container* bileşeninin temel amacı, *Operation* bileşeninde tanımlanan fonksiyonlar için *Data* bileşenini hazırlamaktır.

Operation bileşeni, *Container* üzerinde gerçekleştirilen algoritmalardan sorumludur. Algoritmalar, esneklik ve sürdürülebilirliği artırmak amacıyla veri yapısından (*Container*) ayrılmıştır. Böylece, birleşik yapıyı değiştirmeden, *Operation* sınıfından yeni bir sınıf türeterek kolayca yeni işlemler tanımlanabilir. *Operation* bileşeninin bir diğer amacı ise ham metin verisini yapısal *Data* temsillerine dönüştürerek sistem genelinde bütünlük ve tutarlılığı sağlamaktır.

Lingan'ın kolay kullanımını göstermek için Şekil 2.1'de bir örnek kod parçası verilmiştir. Verilen kodda, *belge* kelimesinin göreceli sıklığını iki farklı zaman diliminde hesaplanır. Bu kod parçasında dolaylı bir varsayım mevcuttur. *Corpus* nesnelerinin oluşturulup lokal diske kaydedildiğini varsaymaktadır. İlk olarak, 1930–1939 ve 1980–1989 yıllarına ait korpuslar yüklenir [Satır 1-2]. Sonrasında, bu korpus objeleri ile birlikte diyakronik bir korpus oluşturulur [Satır 3].

```
1. corpus_1930 = Corpus.load("path/to/corpus_1930")
2. corpus_1980 = Corpus.load("path/to/corpus_1980")
3. dia_corpus = DiachronicCorpus(corpora=[corpus_1930, corpus_1980])
4. frequency_series = dia_corpus.perform(Frequency(word = "belge", normalize=True))
5. print(frequency_series)
```

Şekil 2.1 : *Lingan*'ın örnek bir kullanımı.

Ardından, verilen bir kelimenin göreceli sıklığını hesaplamaktan sorumlu olan yeni bir *Frequency* nesnesi oluşturulur ve diyakronik korpusun *perform* metoduna argüman olarak verilir [Satır 4]. Sonuç, *frequency_series* değişkenine kaydedilir ve ardından yazdırılır [Satır 5].

Araştırmacılar kendi hiyerarşik program yapılarını oluşturmak için *Lingan*'ı kullanabilirken, onlara daha fazla yardımcı olmak amacıyla çeşitli işlevler ve veri

yapıları için de hazır fonksiyonlar ve metotlar sağladık. Özellikle, farklı zaman dilimlerinde kelime dağarcığı benzerliğini hesaplama, belirli bir kelime için verilen bir dönemde en yakın kelimeleri tespit etme, belirli bir zaman aralığı boyunca kullanılan kelimeleri belirleme gibi çeşitli görevler için işlevler geliştirildi. Ek.2 kütüphanede bulunan önemli işlevleri ve açıklamalarını listelemektedir.

2.8 Korpus İstatistiği

Çizelge 2.2 Turkronicles hakkında genel istatistikleri sunmaktadır. Veri seti 45.375 belgeden oluşmakta olup, metin formatında 8.5 GB boyutundadır. Filtreleme adımımızdan önce veri seti yaklaşık 849 milyon token ve 1.961.044 çeşit sözlük

Çizelge 2.2 : Turkronicles korpusunun genel istatistiği.

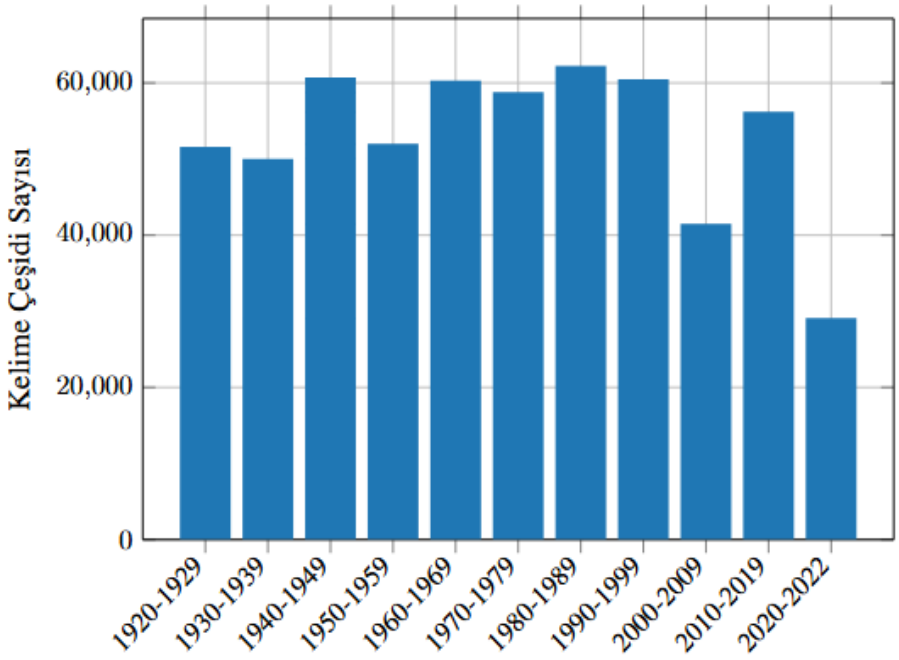
Dokümanların sayısı	45375
Filtrelemeden önce kelime sayısı	849.335.014
Filtrelemeden sonra kelime sayısı	842.957.298
Filtrelemeden önce indirgenmiş kelime çeşidi sayısı	1.961.044
Filtrelemeden sonra indirgenmiş kelime çeşidi sayısı	211.775
Kelime çeşidi sayısı	10.689.405
Doküman başına düşen token sayısı	18.718

formuna indirgenmiş kelime içermektedir. Filtreleme işleminden sonra toplam token sayısı 842.957.298'e, indirgenmiş kelime çeşidi sayısı ise 211.775'e düşmüştür.

2.9 Kelime Dağarcığının Değişimi

Bu bölümde korpusta bulunan dokümanlar 10 yıllık zaman dilimlerine ayrılarak farklı açılardan karşılaştırılmıştır. Tüm hesaplamalarda, sözlük formuna indirgenmiş kelimeleri kullanıldı. İlk olarak, zaman dilimlerine göre kelime dağarcığı büyüklüğünü elde edildi. Şekil 2.2, her bir zaman dilimi için indirgenmiş formdaki kelime çeşidi sayısını göstermektedir. Kelime dağarcığı büyüklüğünün genel olarak zaman dilimleri arasında dengeli olduğunu gözlemlendi. Ancak, 2020-2022 için indirgenmiş kelime çeşidi sayısı, doküman sayısının azlığı nedeniyle diğer tüm zaman dilimlerinden daha düşüktür. 2000-2009 yılları arasındaki kelime dağarcığı büyüklüğü, 2020-2022 haricindeki diğer dönemlerden daha azdır. Bunun nedeni, bu dönem için metin çıkarma sürecinde PDF dosyaları yerine HTML dosyalarının kullanılması olabilir. Aslında bu durumda optik tarama hatalarının daha az olması

beklenti dahilindedir. Çünkü dokümanlar dijital olarak hazırlanmıştır. Dolayısıyla kelime çeşidi sayısı da daha az çıkması beklenmektedir.



Şekil 2.2 : Her 10 yıllık zaman dilimi için indirgenmiş kelime çeşidi sayısı

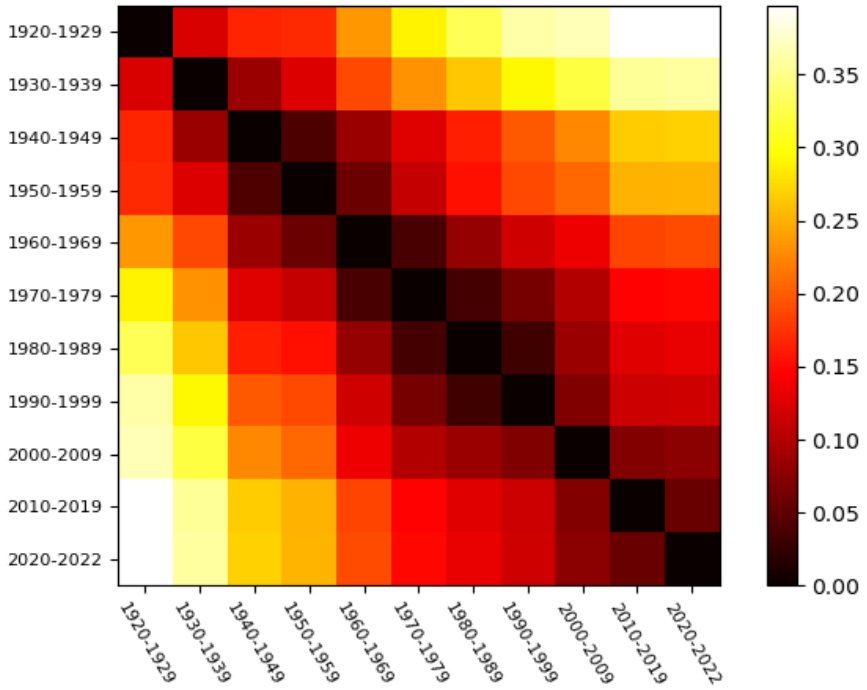
Sıradaki inceleme zaman dilimleri arasındaki kelime dağarcığı farklarına odaklanmaktadır. İlk olarak, her 10 yıllık dönem için kelime çeşitlerinin olduğu birer küme oluşturuldu. Ardından, farklı zaman aralıklarındaki dokümanlar arasında Jaccard Index (JI) ve Jensen-Shannon Sapması (JSD) değerleri hesaplandı. Bu metriklerin matematiksel gösterimleri sırasıyla Eşitlik (2.3-2.4)'te verilmiştir. Eşitlik 2.3'te yer alan V_{T_1} ve V_{T_2}, T_1ve T_2 zaman dilimlerine ait kelime kümesi, Eşitlik 2.4'te yer alan P_{T_1}ve P_{T_2} ise T_1ve T_2 zaman dilimlerindeki unigram olasılık dağılımı temsil etmektedir. Eşitlik 2.4'te bulunan M değişkeni ise ortalama dağılımdır ve $M = \frac{1}{2}\left(P_{T_1} + P_{T_2}\right)$ şeklinde elde edilir.

$$\mathrm{JI}\left(\mathrm{V_{T_1}}, \mathrm{V_{T_2}}\right) = \frac{\left|\mathrm{V_{T_1}} \cap \mathrm{V_{T_2}}\right|}{\left|\mathrm{V_{T_1}} \cup \mathrm{V_{T_2}}\right|} \tag{2.3}$$

$$\mathrm{JSD}\left(\mathrm{P_{T_1}} \parallel \mathrm{P_{T_2}}\right) = \frac{1}{2}\mathrm{KL}\left(\mathrm{P_{T_1}} \parallel \mathrm{M}\right) + \frac{1}{2}\mathrm{KL}\left(\mathrm{P_{T_2}} \parallel \mathrm{M}\right) \tag{2.4}$$

Şekil 2.3, her zaman dilimi ikilisi için JSD değerini göstermektedir. Bu ısı haritasını incelediğimizde zamana dilimleri arasındaki uzaklık arttıkça, farklılaşmanın da

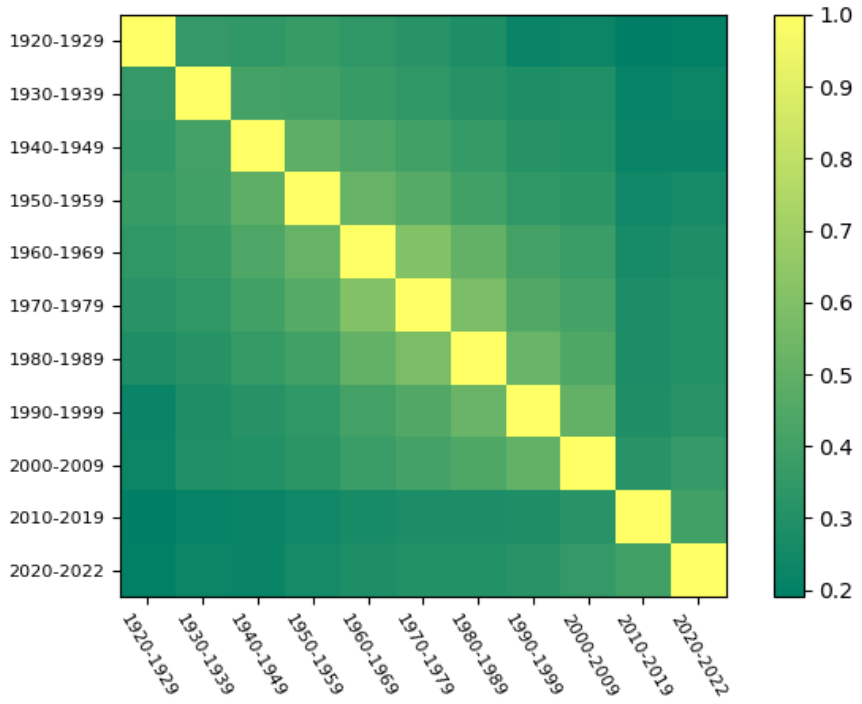
arttığı gözlemlendi. Benzer şekilde, Şekil 2.4 her bir zaman dilimi ikilisi için JI benzerlik değerlerini ısı haritası formatında gösteriyor.



Şekil 2.3: Zaman dilimleri arasındaki JSD değerlerini gösteren harita

İki ardışık zaman dilimi arasındaki JI değerinin yaklaşık yarısında %50'nin altında olduğunu gözlemlendi. Ayrıca, 1920'lerdeki kelime dağarcığı ile 1990'lardaki kelime dağarcığı arasındaki benzerlik yaklaşık %0,2'dir, bu da dilin 100 yıl içinde dramatik bir şekilde değiştiğini göstermektedir. Bu büyük değişimi göstermek için, Çizelge 2.1'de gösterilen ilk cümlenin modern Türkçe karşılığı yeniden yazıldığında "*Egemenlik (Hâkimiyet) kayıtsız şartsız (bilâ-kayd ü şart) milletindir. Yönetim (idare) şekli (usulü), halkın yazgısını (mukadderatını) doğrudan doğruya (bizzat) ve fiilen/gerçekten (bilfiil) yönetmesi (idare etmesi) esasına dayanmaktadır (müstenittir)*" cümlesi ortaya çıkmaktadır. Orijinal cümledeki kelimeler parantez içinde, modern karşılıkları kırmızı renkte, değişmeyen kelimeler ise siyah renkte belirtilmiştir. Orijinal cümlede 15 kelime olup, bunların 11'i (%73,3) modern karşılıklarıyla değiştirilmiştir.

Kelime dağarcığındaki değişimin daha ayrıntılı incelenmesi için, 1930-1939 ve 1980-1989 JSD değerine katkısına göre sıralandı (Pechenick ve diğ., 2015). Şekil 2.5, 1930-1939 ve 1980-1989 yıllarına ait dokümanlar arasındaki JSD değerine en büyük katkı sağlayan 60 kelimeyi göstermektedir.



Şekil 2.4 : Zaman dilimleri arasındaki JI değerlerini gösteren harita

Şekil 2.5'te, 1980-1989 dönemindeki birçok modern Türkçe kelimenin karşılık geldiği Arapça ve Farsça kökenli kelimelerin yerine geçtiği gözlemlendi. Yatay çubuklar, karşılık geldiği kelimenin JSD skoruna katkısının büyüklüğünü göstermektedir. Değerlerin işareti, kelimelerin göreceli olarak daha sık bulunduğu zaman dilimini belirtir; soldaki kırmızı çubuklar 1930-1939 döneminde daha yaygın olan kelimeleri, sağdaki mavi çubuklar ise 1980-1989 döneminde daha sık kullanılan kelimeleri temsil etmektedir. Çizelge 2.3, Şekil 2.5'te farklı zaman dilimlerinde ortaya çıkan ancak aynı veya benzer anlama sahip tüm kelime çiftlerini listelemektedir. Örneğin, hem *reis* hem de *başkan* kelimeleri Türkçe'de aynı anlama gelir; ancak *reis* kelimesi 1930-1939 döneminde sık geçen ve JSD skoruna en çok katkı sağlayan kelimelerden biriyken, *başkan* 1980-1989 döneminin en farklılık gösteren kelimelerinden biridir. Ayrıca, bazı kelimelerin karşılaştırılan dokümanlardaki içerik veya üslup değişiklikleri nedeniyle Şekil 2.5'te ortaya çıktığı gözlemlendi. Örneğin, 1980-1989 dönemindeki Resmî Gazete'de yer alan uluslararası anlaşmalar ve sözleşmeler nedeniyle *the* ve *of* gibi İngilizce kelimeler iki zaman dilimi arasında en çok farklılık yaratan kelimeler listesinde yer almaktadır. Üslup değişikliğine ilişkin olarak ise, *Türk Lirası*'nın kısaltması olan *TL* ve *lira*, farklı zaman dilimlerinden gelen iki ayrı terim olarak rastlanmaktadır; bu durum, *lira*

için *TL* kısaltmasının benimsenmesine yönelik bir kaymayı işaret etmektedir. Arapça ve Farsça kökenli kelimelerin yerini almak üzere nasıl yeni kelimelerin ortaya çıktığını daha iyi anlamak için aynı anlama sahip iki farklı kelime ikilisine odaklanıldı: i) *gerek* ve *mucip* ve ii) *yıl* ve *sene*.

Çizelge 1.3 : Benzer veya aynı anlama sahip olup farklı zaman dilimlerinde JSD skoruna en çok katkı sağlayan bazı kelimeler.

1930-1939	1980-1989
vekil	bakan
sene	yıl
umumi	genel
reis	başkan
heyet, encümen	kurul
vesika	belge
icra	uygula
mucip	gerek
aza	üye
idare (et)	yönet
sayı	numara
layiha	tasarı

Şekil 2.6 ve Şekil 2.7, bu kelimelerin normalize edilmiş frekanslarının yıllara göre değişimlerini göstermektedir. *gerek* ve *mucip* 1920'lerde mevcut olmasına rağmen, Arapça kökenli *mucip, gerek* kelimesinden daha sık kullanılmaktadır. Ancak sonraki dönemlerde gerek kelimesi *mucip*'e göre daha popüler hale gelmiştir ve *mucip* kelimesi 1970'lerden sonra dokümanlarda gözükmek. İkinci örnekte, başka bir ilginç durum gözlemlendi. *yıl* kelimesi, korpusta 1920'lerde mevcut değildir; ancak 1930'lardan sonra *sene* kelimesinden daha yaygın hale gelmiştir.

2.10 Zaman Periyotları Boyunca Semantik Benzerlik

Bu bölümde kelime vektörleri ile diyakronik bir analiz yaparak, farklı zaman dilimlerinde anlamsal olarak benzer kelimelerin tespit edilip edilemeyeceği araştırıldı. JSD analiziyle tutarlılık sağlamak açısından aynı zaman aralıkları 1930-1939 ve 1980-1989 dönemleri olarak belirlendi. Daha sonra, 1980-1989 döneminin kelime vektörleri, 1930-1939 döneminin vektör uzayına Ortogonal Procrustes yöntemi kullanılarak hizalandı. Kelime çiftleri arasındaki benzerliği ölçmek için kosinüs benzerliği kullanıldı. İlk olarak, 1980-1989 döneminde JSD analizimize göre en çok farklılık gösteren 12 kelime üzerinden ve her kelime için 1930-1939 dönemindeki en yakın 10 kelime belirlendi. Bulunan en yakın 10 kelimenin içinde,

ilk kelime veya ikinci kelime, çoğunluk durumda 1980'ler dönemindeki karşılık gelen kelimeyle aynı anlama sahiptir.

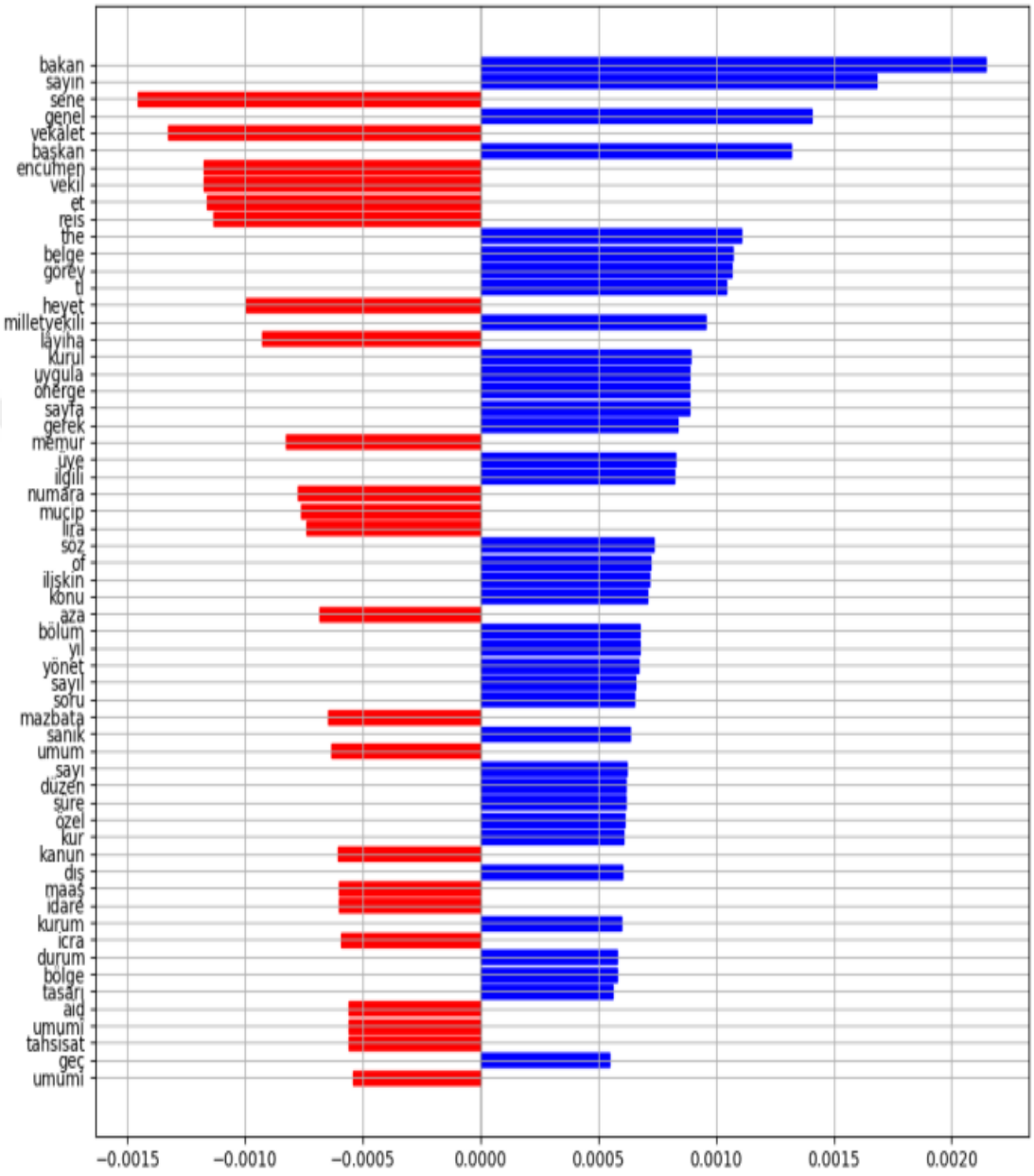
Çizelge 2.4 : 1980'lerde kullanılan kelimeler ilk sütunda, bu kelimelere 1930'larda kullanılan en benzer on kelime ikinci sütunda gösterilmektedir.

1980-1989	1930-1939 yılları arasındaki 10 en yakın kelime
bakan	**vekâlet**, *iktıs*, *îktıs*, *vekâl*, içtimaî, maarif, *tktıs*, iktisat, nafıa, îcra
yıl	**sene**, **yıl**, **yılma**, ayı, takvim, aylık, seneye, iptida, *dörd*, katıl
belge	**vesika**, ibraz, istek, *vesai*, makbu, musaddak, mütea, **vesaik**, talih, makbuz
gerek	icabe, **iktiza**, göre, **icap**, ayrıç, kanunî, **zarurî**, kabîl, **lüzum**, önce
başkan	**reis**, müteşekkil, seçim, *zatte*, müsteşar, inha, vekâlet, mütal, **riyaset**, seç
genel	müdür, **umum**, işle, **umumî**, idare, *teşki*, denizyolu, **genel**, havayolu, îdare
kurul	seçim, **heyet**, îcra, ödev, baro, seçilir, yönetim, inha, teşekkül, seç
uygula	**tatbik**, göre, gözet, *dışmd*, hüküm, önce, **tatbikat**, *icabl*, istisnaî, *tatbi*
üye	seç, seçilir, seçilmiş, seçim, **üye**, intihap, seçi, *zatte*, müntahap, müntehap
yönet	talimatname, talimat, nizamname, teşkilat, bölüm, teşki, izahname, sayıl, ilgili, *plânl*
numara	**numara**, yazı, değiştiri, aşağı, ilişik, gösteri, **sayı**, mezkûr, ün, yaz
tasarı	*lâyih*, **lâyiha**, *eneüm*, *encüm*, değişik, mütenazır, tadil, bazı, encümen, *ncü*

Bu duruma uymayan *üye* ve *yönet* kelimeleri için bulunan kelimelerin neredeyse tamamı semantik olarak ilişkili kelimelerdir. İlginç bir şekilde, *üye* kelimesi her iki dönemde de yer almakta, ancak 1980'lerdeki *üye* kelimesini 1930'ların uzayını hizalanmış vektörüne en yakın beşinci kelime olarak yer almıştır. Çizelge 2.4 sonuçları göstermektedir.

Turkronicles korpusu hizalama yöntemleriyle birlikte kullanıldığında belirli bir zaman diliminde mevcut olmayan bir kavram ile ilgili kelimeleri tespit etmeyi de mümkün kılmaktadır. Örneğin, *helikopter* kelimesi 1930-1939 kelime dağarcığında bulunmamaktadır ve o yıllarda bilinen bir kavram değildi. 1930'larda helikopter ile ilişkili kelimeleri incelemek için, 1980-1989 dönemindeki *helikopter* kelimesinin vektörünü 1930-1939 vektör uzayına hizaladık. Ardından 1930-1939 yılları arasında *helikopter* kelimesine en yakın olan on kelimenin ulaşım kavramıyla yakından ilişkili olduğunu bulduk: *nakliyat*, *vasıta*, *hava*, *motor*, *deniz*, *otobüs*, *motör*, *binek*, *sefine* ve *gemi*.

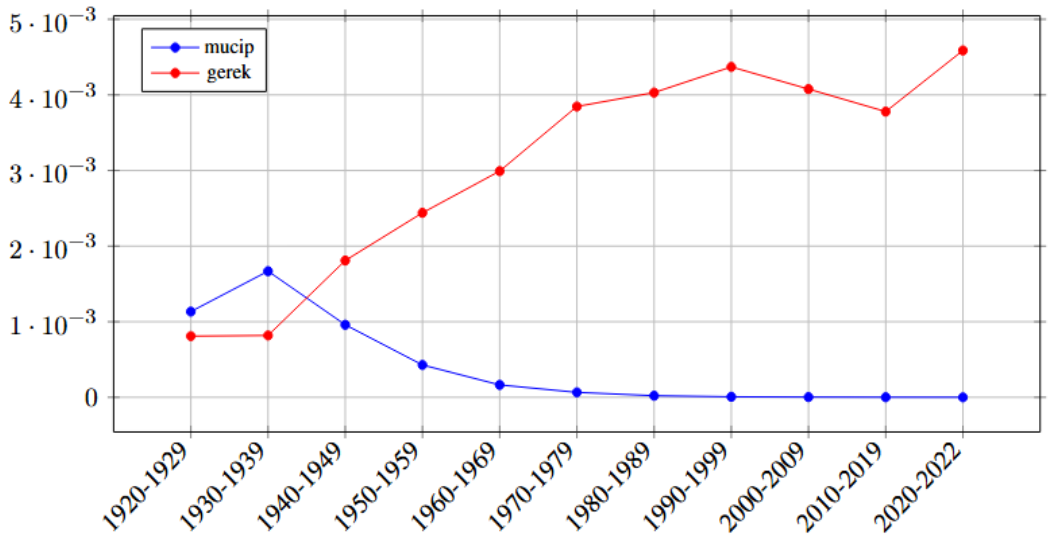
Analizler bütünüyle ele alındığında, Turkronicles korpusu üzerinden oluşturulan kelime vektörlerinin, farklı zaman dilimlerinde kullanılan kelimeler arasındaki anlamsal ilişkileri keşfetmeyi ve incelemeyi mümkün kıldığını ortaya koymaktadır.



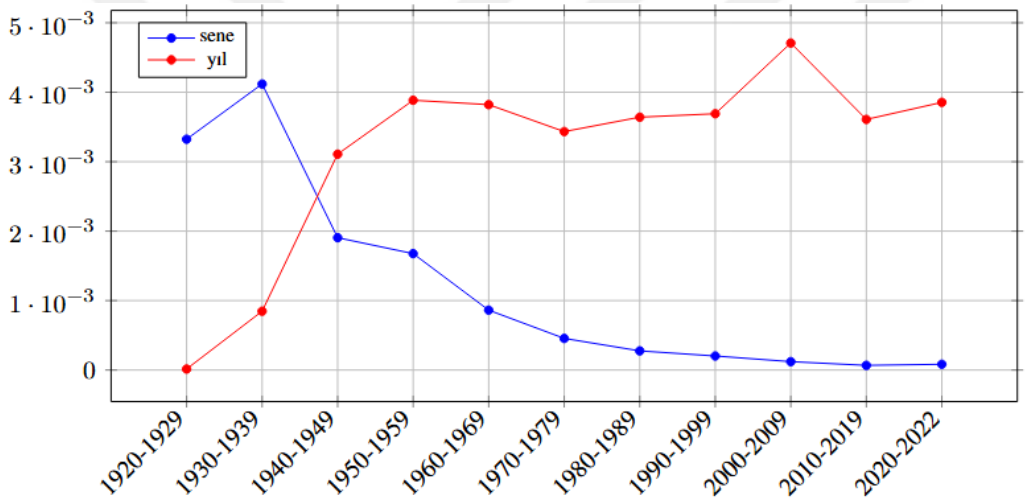
Şekil 2.5: 1930-1939 ve 1980-1989 dönemlerinden Jensen-Shannon farkına katkılarına göre sıralanan ilk 60 kelime.

2.11 Yazım Kurallarında Değişim

Yazım kurallarındaki değişiklikleri incelemek için iki farklı analiz gerçekleştirildi. İlk olarak, kelimelerin sonlarının nasıl değiştiği ve ardından düzeltme işaretinin kullanım sıklığının zamana göre değişimi incelendi. Türkçe kökenli kelimeler, birkaç



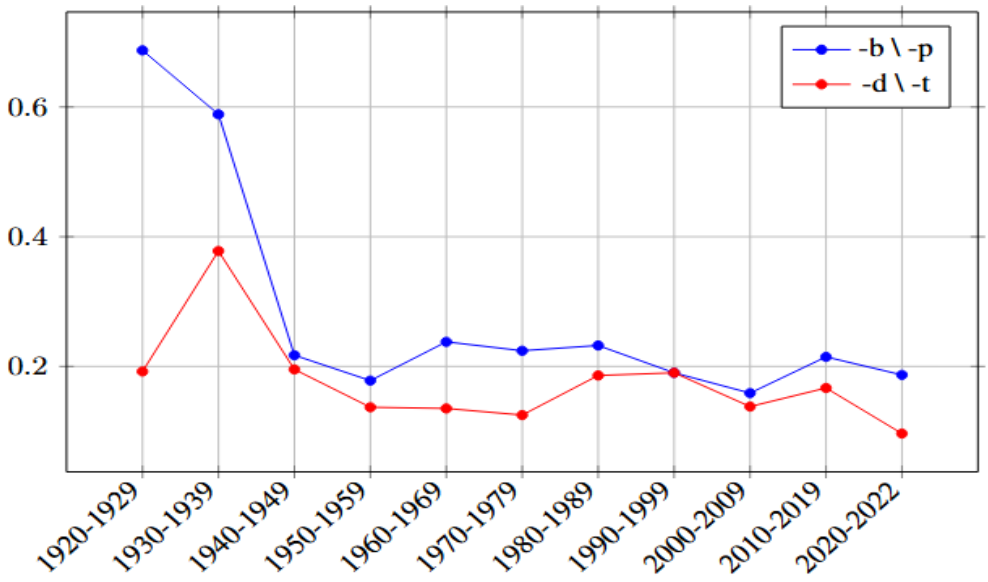
Şekil 2.6 : *mucip* ve *gerek* kelimelerinin yıllara göre frekans değişimi



Şekil 2.7 : *sene* ve *yıl* kelimelerinin frekansının yıllara göre değişimi

istisna dışında *b, c, d* ve *g* harfleriyle bitmez. Ancak, Arapça ve Farsçadan alınan ve bu harflerden biriyle biten birçok ödünç kelime bulunmaktadır.

Bu kelimelerin, son harfin *b,c,d,g* yerine *p,ç,t,k* olarak değiştiği iki farklı şekilde yazıldığı gözlemlenmiştir (örneğin, Ahmet vs. Ahmed). Bu kelimelerin dönüşümünü gözlemlemek için, önce bu fonetik kurala göre tek bir harfin farklı olduğu kelimeler tespit edildi. Daha sonra her 10 yıllık zaman dilimi için *d* ve *b* ile biten kelimeler ile bunların *t* ve *p* ile biten karşılıklarının frekanslarını hesapladık. Nadiren olmaları sebebiyle *c-ç ve g-ğ* harfleriyle biten kelimeler hesaplama aşamasına dahil edilmedi. Son olarak, *b/p* harfleriyle biten kelimeler, *d/t* harfleriyle biten kelimelere oranı hesaplandı. Sonuçlar Şekil 2.8'de gösterilmektedir.



Şekil 2.8 : Farklı zaman dilimlerindeki, "-b/-d" harfleriyle biten kelimelerin "-p/-t" harfleriyle biten kelimelere oranı.

Şekil 2.8 incelendiğinde her iki harf çifti için de oran 1'in altındadır; bu, tüm zaman dilimlerinde kelimelerin *t* ve *p* harfleriyle bitmesinin, *d* ve *b* harfleriyle bitmesinden daha yaygın olduğunu göstermektedir. Ancak, bu durum kullandığımız morfolojik analiz aracından kaynaklanıyor olabilir. Özellikle, araç modern Türkçe dilbilgisi kurallarına dayalı olarak geliştirildiğinden, bazı durumlarda kökleri *t* ve *p* ile bitiyormuş gibi tanımlayabilir. Bu nedenle, gerçek değerler yerine eğilime odaklanmak önemlidir. *b/p* oranı 1920'ler ile 1940'lar arasında önemli ölçüde azalmaktadır. Daha sonra, *b* harfiyle biten kelimelerin yüzdesi 0.16 ile 0.24 seviyeleri arasında dalgalanmaktadır. Ancak, *d* harfiyle biten kelimeler için farklı bir örüntü gözlemliyoruz. İlginç bir şekilde, *d* harfiyle biten kelimelerin yüzdesi önce artmakta (1920'lerden 1930'lara) ve ardından azalmaktadır. Sonraki zaman dilimlerinde, *d/t* oranı 0.1 ile 0.2 arasında dalgalanmaktadır. Bu dalgalanmalar, korpusun yetersizliğinden veya insanların dildeki reformlara karşı direnç göstermesinden kaynaklanıyor olabilir.

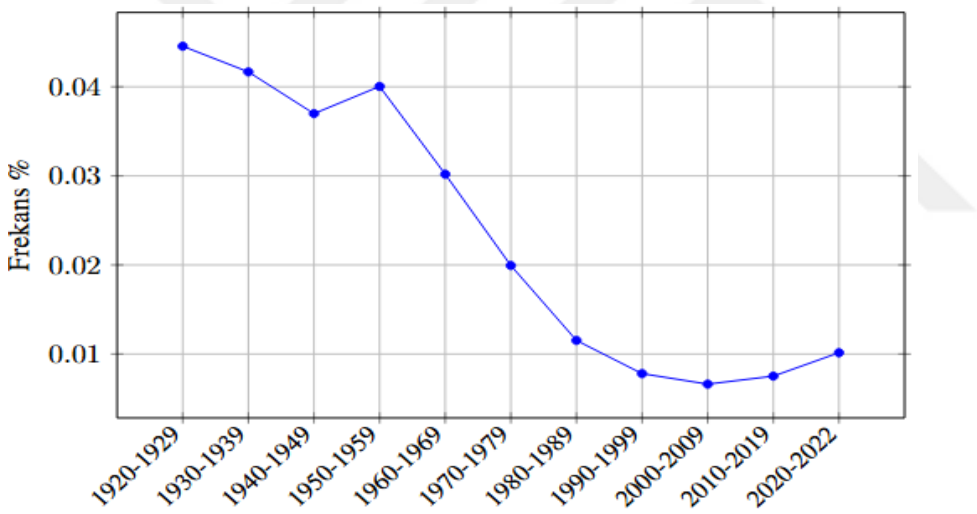
Türkçe'deki imla değişikliklerileriyle ilgili ilginç bir konu, şapka işareti (^) taşıyan harflerin kaldırılmasına dair şehir efsanesidir. Özellikle bazı kelimelerde *a*, *ı* ve *u* gibi harfler şapka işaretiyle yazılır, örneğin, *kâğıt*, *abidevî, şûra* vb. Bu harfler kaldırılmamış olmasına rağmen, birçok kişi sosyal medya platformlarında önce kaldırıldığı, ancak daha sonra geri getirildiği iddiasında bulunmaktadır. Hatta

doğruluk kontrolü yapan web siteleri bile bu iddianın gerçekliğini doğrulamak zorunda kalmıştır[11].

Turkronicles korpusunda, bu şehir efsanesine odaklanarak düzeltme işareti taşıyan harflerin sıklığının değişimi incelendi. Sonuçlar Şekil 2.9'da gösterilmektedir. Düzeltme işareti taşıyan harflerin zaman içinde değişen sıklıklarda kullanıldığını, ancak kullanılmaya devam edildiği gözlemlenmektedir. Bununla birlikte, bu harflerin sıklığının önemli ölçüde azaldığı da gözlemlenmektedir.

2.12 Sınırlamalar

Çalışmamız, 1920'lerden bu yana Türk dilinin nasıl değiştiğine dair değerli içgörüler sunsa da Turkronicles kullanılırken dikkate alınması gereken bazı hususlar bulunmaktadır. Turkronicles, ağırlıklı olarak yönetime ilişkin konularda yetkililer tarafından kullanılan politik dilin kullanıldığı dokümanları içermektedir.



Şekil 2.9 : *Turkronicles*'ta şapka işaretinin kullanım sıklığı.

Bu nedenle, Türkçe'nin tüm özelliklerini kapsamlı bir şekilde temsil etmemektedir. Bununla beraber, Türkçe'nin nispeten hızlı değişim geçirmesinin başlıca nedenlerinden biri, Arapça veya Farsça kökenli kelimelerin yerine yeni kelimeler önermek ve alfabeyi değiştirmek gibi dil reformunu hedefleyen politikalardır. Bu bağlamda, Turkronicles, Türk dil reformu üzerinde alınan kararların etkisini araştırmak için en iyi kaynaklardan biri olabilir.

[11] https://www.malumatfurus.org/sapka-isaretinin-kaldirildigi-iddias

Çalışmamızda, PDF dosyalarından metin çıkarma ve kök bulma işlemlerini otomatikleştirmek için araçlar kullandık. Ancak, ortaya çıkan sonuçlar kullandığımız araçların başarısına ve diğer etkenlere doğrudan bağlıdır. Örneğin, analizlerde görülen bazı kelimelerin hiçbir anlam taşımadığını gözlemliyoruz (Bkz. Çizelge 2.4). Yazım hatası içeren bu kelimeler, eğitim aşamasında kelime vektör modellerini de yanıltmaktadır. Bu nedenle, aşırı nadir kelimeleri filtrelemek gibi gürültüyü azaltmaya yönelik önlemler alıyoruz. Bu sorunu daha da hafifletmek için, bulgularımızın tekrarlanabilirliğini artırmak ve korpusumuz üzerinde daha fazla araştırma yapılmasını sağlamak amacıyla kodumuzu ve ham verilerimizi de paylaşıyoruz. Yine de bulgularımızı analiz ederken kullandığımız araçların olası etkileri dikkate alınmalıdır.

3. FARKLI ZAMAN DİLİMLERİNDE KULLANILAN EŞ ANLAMLI KELİMELERİN TESPİTİ

Dil, insanların anlam iletmek için kullandığı bir iletişim kanalıdır. Dillerin en belirgin özelliklerinden biri dinamik yapılarıdır. Bir dili oluşturan temel bileşenler zamanla değişebilir (Aitchison, 2005; Bybee, 2015; McMahon, 1994). Bu değişimler, dilin içinde bulunduğu toplum tarafından şekillenir ve sosyo-kültürel özellikleri yansıtır. 1920'lerin sonlarında başlayan Türk Dil Reformu, Türkçe'yi sosyo-kültürel ve sosyo-politik eylemlerin dil evrimini nispeten kısa bir süre içinde etkilerinin kolaylıkla incelenebileceği bir dil haline getirdi. Türk Dil Reformu ile, ilk adım, 1928'de Latin alfabesine geçildi. Ardından, Osmanlı döneminde sayıca fazla olan Farsça ve Arapça kökenli kelimelerin, Türkçe kökenli kelimelerle değiştirildi (Lewis, 1999).

Bu tarz dildeki değişiklikler tarihî belgelerin analiz edilmesini ve anlaşılmasını zorlaştırmaktadır. Türk Dil Reformu sırasında yapılan kapsamlı çalışmaları, bir dilde bir kavram için yeni bir kelimenin (Fischer, 1998) dilde türemesi, özellikle genç nesillerin cumhuriyetin ilk dönemlerinde ortaya çıkan dokümanları anlamasını zorlaştırmaktadır. Bununla beraber, sözlükler de dilin değişiminden etkilenir: sözlükler bir dilin belirli bir zaman aralığında sahip olduğu kelime varlığının tümünü içeremez. Bu durum onları doğası gereği eksik kılar. Örneğin *müteşir*[12], *bab*[13] ve *vecahi*[14] gibi kelimeler TDK'nin Güncel Türkçe Sözlüğü'nde[15] yer almamaktadır.

Bunlara ek olarak bir dildeki değişiklikler tarihî belgeler üzerinde NLP modellerinin performansını düşürür. Örneğin Büyük Dil Model'lerinin (LLM) performansı, eğitim verisinin zaman diliminden farklı bir dönemden gelen kelimeler veya ifadelerle karşılaştıklarında, zaman uyumsuzluğu durumunda (Luu ve diğ., 2022), önemli ölçüde düşer. Bu durum, NER (Luu ve diğ., 2022; Onoe ve diğ., 2022), ve QA

[12] TBMM Tutanakları – 06.12.1961 – 18. Birleşim
[13] Resmî Gazete – Sayı: 3027
[14] Resmî Gazete – Sayı: 1100
[15] https://sozluk.gov.tr/

(Zheng ve diğ., 2024) gibi görevlerde modellerin doğruluğunu ve güvenilirliğini azaltabilir. Yakın zamanda Zheng ve diğ. (2024), modele verilen metin kesitlerinde neolojizm kelimeler yer aldığında MT'de dramatik bir performans düşüşü gözlemlendiğini bildirmiştir. Bu nedenle, tarihî bir belgede verilen bir kelimenin anlamını tespit edebilecek ve bu belgeleri etkili bir şekilde analiz edebilecek modellere ihtiyaç duyulmaktadır.

Bu bölümde, *Turkronicles* korpusu kullanılarak bir zaman diliminde olan bir kelimenin, farklı bir zaman diliminde anlam olarak yakın veya benzer kelimeleri tespit etmek için iki farklı yöntem önerildi ve literatürde farklı zaman dilimlerindeki analog kelimelerin tespitinde kullanılan bir metot ile önerilen yöntemler karşılaştırıldı. Önerilen yöntemler Ortogonal Procrustes Hizalaması (*OP*) ve Ortogonal Procrustes ile Spearman Sıralama Korelasyonu (*OP+SC*)'dur. *OP*, iki farklı zaman diliminin vektör uzayları arasında dönüşüm matrisiyle uzayların birbirine hizalanmasına dayanır. *OP+SC* ise *OP* metodunun üzerine bir eklenti olarak, bir sıralama mekanizması koyar: kelime ikililerinin belirtilen zaman aralıkları boyunca değişen kullanım sıklıklarını göz önüne alarak Spearman korelasyon katsayısını hesaplayıp küçükten büyüğe sıralar. Deneylerde önerilen yöntemlerin performansını CBOW ve SVD vektörleri üzerinde nasıl değiştiği incelendi. Sorgu zaman dilimi ve hedef zaman dilimleri arasındaki mesafenin artmasıyla bu yöntemlerin performansının nasıl değiştiği de gözlemlendi. Ayrıca Spearman korelasyon katsayısının dahil edilmesi, çoğu durumda hem CBOW hem de SVD vektörlerinde *OP* modelinin performansı arttığı görüldü. Son olarak, önerilen yöntemlerin performansının, hedef zaman dilimi 1960'lardan 1980'lere değiştirildiğinde benzer kaldığını, ancak sonraki zaman dilimlerinde kademeli olarak azalma eğiliminde olduğu tespit edildi.

Literatür araştırmasında, bu çalışma doğal dil işleme yöntemleriyle Türkçe'de büyük ölçekli bir korpus kullanılarak hesaplamalı bir biçimde neolojizme odaklanan ilk çalışma olduğu görüldü.

3.1 Problem Tanımı

Amaç iki farklı dönemde yazılmış belgelerde kullanılan eşanlamlıları belirlemektir. Bu problemi bir sıralama problemi olarak ele almak makuldür. Sorgu zaman dilimi T_Q ve hedef zaman dilimi T_T olsun. T_Q sorgu zaman diliminde kullanılan bir kelime

w için, hedef zaman dilimi T_T'deki kullanılan ve w ile anlam olarak aynı veya yakın kelimelerin bulunması problemin özünü oluşturur. Farklı bir perspektiften bakıldığında bu problem bir sıralama problemidir: T_T'de olan kelimelerin w'ya anlam açısından benzerliklerine göre sıralanmasıdır.

Fakat bu problemin, yalnızca vektör uzaylarındaki benzer kelimeleri bulmanın ötesine geçtiğini belirtmek önem arz eder. İlk olarak, bir kelimenin anlamı zamanla değişebilir. Dolayısıyla, bir kelime her iki zaman diliminde de mevcut olsa bile, anlamları her dönemde farklı olabilir. Bu durum da bir kelimenin iki farklı zaman dilimine ait vektörlerinin aynı olmayacağını gösterir. Ayrıca, bir kelime anlamsal bir değişim geçirmese bile, kelime vektör algoritmalarının işleyiş biçimlerinden ötürü yan etki olarak ortaya çıkan dönüşümler nedeniyle farklı zaman dilimlerinden elde edilen temsilleri doğrudan karşılaştırılamaz. Çünkü karşılaştırılan vektörlerin ait olduğu uzayların eksenleri farklıdır.

3.2 Önerilen Metotlar

3.2.1 Ortogonal procrustes hizalaması (OP)

Sezgisel olarak, bir kelimenin yerine yeni başka bir kelime dile girdiğinde bu yeni kelimenin en azından var olan kelime kadar anlam bilgisini taşıması gerekir. Bu nedenle, anlamı değiştirmeden, yeni kelimeyi eski kelimenin bağlamında kullanılabilinir. Bu sezgiden ve kelime vektörlerinin anlam-bağlam bilgisini iyi bir şekilde ifade ettiği varsayılarak, ilk olarak T_q ve T_t zaman dilimleri için ayrı ayrı kelime vektör uzayları elde edilir. T_q ve T_t zaman dilimlerine karşılık gelen kelime vektör uzayları için sırasıyla ve $\mathbf{W_q}$ ve $\mathbf{W_t}$ gösterimleri kullanılacaktır. Bu aşamadan sonra T_q'dan seçilen bir w kelimesine T_t zaman diliminde anlam olarak en yakın kelimeler bulunur. Anlam olarak yakın kelimeler bulunurken w kelimesinin vektör gösteriminin, $\mathbf{w_t} \in \mathbf{W_t}$, T_t zaman diliminde bulunması gereklidir. Ancak bu kelime hedef zaman periyodunda olmayabilir ve eğer yoksa diğer kelimeler ile böyle bir anlamsal kıyaslama yapılamaz. Bu nedenle ilk olarak w kelimesinin vektör gösterimini T_t zaman diliminde oluşturmak gereklidir. Bu noktada ilk çözüm olarak w kelimesinin T_q zaman dilimindeki vektörünün $\mathbf{w_q} \in \mathbf{W_q}$ doğrudan kullanılması düşünülebilir. Fakat farklı zaman dilimlerinden vektörler iki farklı uzayın vektörleri

olacağı için bir uzayın vektörü başka bir uzayda doğrudan kullanılamaz (Kulkarni ve diğ., 2015; Hamilton ve diğ., 2016). Bu nedenle, vektörler arasında doğru bir karşılaştırma yapılabilmesi için iki vektör uzayı birbiriyle hizalanmalıdır. Bu hizalama işlemi Ortogonal Procrustes yöntemiyle yapılır. Ortogonal Procrustes iki vektör uzayını hizalarken ortogonal dönüşüm kısıtı sebebiyle vektörler arasındaki açıları korumaya çalışır. Hizalama işlemin sonunda iki vektör uzayı arasında bir ortogonal dönüşüm matrisi $\boldsymbol{Q}$ elde edilir. Bu dönüşüm matrisini kullanarak w kelimesinin T_t zaman dilimindeki vektörü inşa edilir.

Bu ihtiyaçlar doğrultusunda, $\mathbf{W}_{\mathrm{q}}$'yu $\mathbf{W}_{\mathrm{t}}$'ye hizalamak için *Ortogonal Procrustes* yöntemi kullanılmıştır. Yöntemin objektif fonksiyonu şu şekilde formüle edilir:

$$argmin||\mathbf{E_q Q} - \mathbf{E_t}||_F \tag{3.1}$$

Burada $\mathbf{Q} \in R^{d \times d}$ T_q ve T_t zaman dilimleri arasında köprü görevi görür ve bir zaman diliminden diğerine anlam açısından geçişi sağlar. d vektör uzaylarının boyutudur. $\mathbf{E}_{\mathrm{q}}$ ve $\mathbf{E}_{\mathrm{t}}$, T_q ve T_t zaman dilimlerinde ortak bulunan kelimelerin indeks olarak hizalanmış vektör matrisleridir. Daha açık bir şekilde $\mathbf{E}_{\mathrm{q}}$ ve $\mathbf{E}_{\mathrm{t}}$ aşağıdaki biçimdedir:

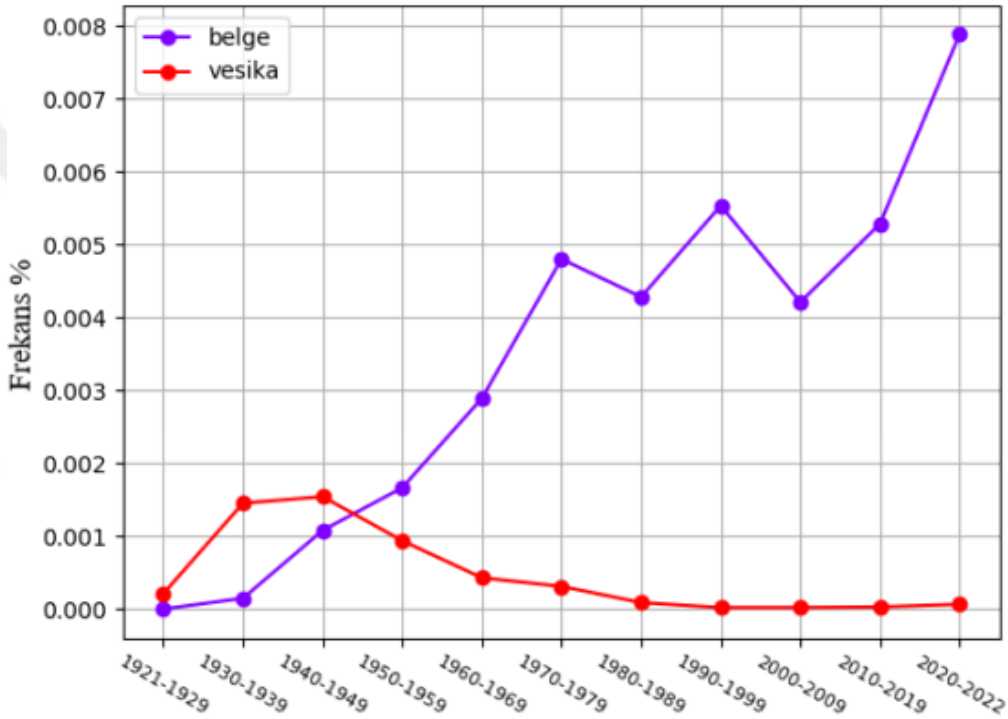
$$\mathbf{E_q} = \begin{bmatrix} -\mathbf{w_1^q} - \\ \vdots \\ -\mathbf{w_s^q} - \end{bmatrix} \qquad \mathbf{E_t} = \begin{bmatrix} -\mathbf{w_1^t} - \\ \vdots \\ -\mathbf{w_s^t} - \end{bmatrix}$$

İki zaman dilimi arasındaki ortak kelime sayısının s olduğunu varsayılsın. Bu durumda $\mathbf{w}_{\mathbf{i}}^{\mathrm{j}}$ i'nci kelimenin j'nci zaman dilimindeki vektörel gösterimidir ve $0 \leq i \leq s, j \in \mathrm{q,t}$. R'nin optimum değeri, $\mathbf{M} = \mathbf{W}_{\mathbf{q}}^{\mathrm{T}}\mathbf{W}_{\mathbf{t}}$ matrisine SVD uygulanarak bulunabilir (Schönemann, 1966).

$\mathbf{Q}$ matrisini bulduktan sonra, w kelimesinin vektörünü $\mathbf{w}_{\mathrm{q}} \in \mathbf{W}_{\mathrm{q}}$, $\mathbf{w}' = \mathbf{Q}\mathbf{w}_{\mathrm{q}}$ dönüşümü ile hizalayarak w'nun T_t zaman dilimindeki vektörel gösterimi $\mathbf{w}'$ oluşturulur. Burada yeni oluşturulan $\mathbf{w}'$ vektörünün, w kelimesinin aranan karşılığıyla hedef zaman diliminde aynı semantik alana karşılık gelmesi beklenir. Son olarak, $\mathbf{w}'$ vektörüne anlamsal olarak benzer kelimeleri k-komşuluğa bakarak kosinüs benzerliğine göre sıralanır ve aranılan kelime tespit edilir.

3.2.2 Ortogonal procrustes hizalaması ile Spearman katsayısı (OP+SC)

1920'lerin ikinci yarısından itibaren, diğer dillerden ödünç alınan kelimelerin yerine Türkçe kökenli birçok yeni kelime dile dahil edilmiştir. Dolayısıyla, ödünç kelimelerin kullanımının zamanla azalması, buna karşılık yeni tanıtılan kelimelerin sıklığının artması beklenebilir. Örnek olarak, *belge* kelimesi, Arapça kökenli *vesika* kelimesinin yerine kullanılmak üzere tanıtılan yeni bir Türkçe kelimedir. *Turkronicles* veri setinde, bu kelimelerin 1920'lerden 2020'lere kadar olan sıklıklarındaki değişim Şekil 3.1'de gösterilmektedir.



Şekil 3.1 : 1920 ile 2022 yılları arasında Turkronicles veri setindeki on yıllık dönemler için *belge* ve *vesika* kelimelerinin göreceli sıklığı.

Bu gözlem OP'nin performansını artırmak için hesaplama aşamasına dahil edilir. İlk olarak, OP'yi kullanarak sorgu kelimesi w için T_b zaman diliminde k en yakın komşu kelimelerin oluşturduğu K kümesi elde edilir. Daha sonra, her aday ikili $(w, c), c \in K$, için bir frekans-zaman serisi oluşturulur. Ardından, bu ikilileri Spearman Korelasyon Katsayısına göre artan sırayla yeniden sıralanır. Bu yöntem, w kelimesinin K kümesindeki aranan karşılığının sıralamasını yükseltir.

3.3 Deneyler

3.3.1 Önişleme

Turkronicles veri kümesini oluştururken uygulanan önişleme aşamaları değişiklik yapılmadan aynı şekilde uygulandı.

3.3.2 Test Kümesi

Önerdiğimiz yöntemlerin performansını değerlendirmek için manuel olarak bir veri seti oluşturduk. Yeni kelimelerin eski karşılıklarının kapsamlı bir listesini belirlemek oldukça yoğun bir efor gerektiğinden, aşağıdaki strateji benimsendi. İlk olarak, sorgu zaman dilimini $T_q = 1980 - 1989$, hedef zaman dilimi ise $T_t = 1930 - 1939$ olarak seçildi. Daha sonra, bu seçilen zaman aralıklarında geçen kelimeler için unigram dağılımları oluşturuldu. Son olarak, bu unigram dağılımları kullanılarak, JSD değeri hesaplandı. Bu hesaplamada, JSD skoruna en fazla katkıda bulunan ilk 5.000 kelime seçildi. Daha sonra bu kelimeler, ilgili zaman dilimlerindeki göreceli frekanslarına göre kategorize edildi. Bu süreç, frekansı bu dönemler arasında önemli ölçüde değişen ve potansiyel olarak yerini başka kelimelere bırakmış kelimelerin belirlenmesini sağladı. Ardından, Güncel Türkçe Sözlük kullanılarak, bu 5.000 kelime arasından sorgu döneminde yer alan kelimeleri (w) ve onun hedef dönemindeki eski karşılığını (c) içeren kelime ikilileri (w, c) manuel olarak oluşturuldu. Sonuç olarak, 221 kelime çifti içeren bir test kümesi elde edildi.

3.3.3 Deney Parametreleri

Deneylerde, farklı parametre değerlerinin yöntemlere farklı etkisini incelendi. Bölüm 2'de belirtilen SVD ve CBOW algoritmaları sonucunda oluşturulmuş kelime vektörleri kullanıldı. Kelime vektörleri 1921'den 2022'ye kadar olan her 10 yıllık dönem için elde edilmiştir.

3.3.4 Karşılaştırılan Yöntem

Önerilen yöntemler, temel yöntem olarak Zhang ve diğ., (2015) 'in doğrusal dönüşüm (LT) yaklaşımı ile karşılaştırıldı. LT ile kelimelerin farklı zamanlardaki analogları bulunurken önemli olan nokta hizalanan zaman dilimlerine ait vektör

uzayları arasında bir doğrusal dönüşüm matrisi oluşturmaktır. Dönüşüm matrisi şu Eşitlik 3.2'deki gibi elde edilir:

$$argmin \sum_{x \in K} || \mathbf{M}\mathbf{x_q} - \mathbf{x_t}||^2 + \alpha||\mathbf{M}||^2 \tag{3.2}$$

Burada **M** doğrusal dönüşüm matrisi, K her iki dönemde de ortak bulunan kelimeleri içeren küme, $\mathbf{x_q}$ ve $\mathbf{x_t}$ bu ortak kelimelerin sırasıyla sorgu ve hedef dönemlere ait vektörleridir. Zhang ve diğ., (2015)'te, kelime çiftleri hem sorgu hem de hedef zaman dilimlerinde ortak olan ve frekansa göre ilk %5'lik dilimde bulunan kelimeler olarak tanımlanmıştır. Deneylerimizde, LT için ortak kelimeler her iki dönemde de en sık geçen 1.000 kelime arasından seçildi. Son olarak, $\alpha = 0.2$ parametresiyle Ridge Regresyon modeli eğitilerek **M** matrisini elde edildi.

3.3.5 Değerlendirme Metrikleri

Problem, her sorgu kelimesi için yalnızca bir doğru kelimenin bulunduğu bir sıralama problemi olarak ele alındı. Bu nedenle, yöntemlerin performansını ölçmek için bilgi erişim alanından Duyarlılık (Recall) ve Ortalama Karşılıklı Sıra (MRR) metrikleri kullanıldı. Bu metriklerden MRR, sorgu için yalnızca tek bir doğru karşılık olduğunda etkili olan bir yöntemdir ve Eşitlik 3.3'te verildiği şekilde hesaplanabilir:

$$MRR = \frac{1}{|N|} \sum_{i=1}^{|N|} \frac{1}{rank_i} \tag{3.3}$$

Burada $|N|$ sorguların (kelimelerin) sayısını, $rank_i$ i-inci sorgu için ilk doğru kelimenin sıralama pozisyonunu ifade eder.

Duyarlılık metriği sistemin doğru öğeleri ilk k sonuç içinde ne kadar iyi elde ettiğini ölçmek için kullanılır ve Eşitlik 3.4'teki gibi formüle edilir.

$$Recall@k = \frac{TP_k}{TP_k + FN_k} \tag{3.4}$$

Burada TP_k ilk k sonuç içindeki doğru pozitiflerin sayısını; FN_k, ilk k sonuç içindeki yanlış negatiflerin sayısını ifade eder. Deneylerde, k değerleri 1, 10 ve 100

olarak belirlendi. OP+SC yöntemi için, sırasıyla Recall@1, Recall@10 ve Recall@100 metrikleride en yakın 5, 15 ve 150 komşuyu dikkate alıyoruz.

3.3.6 Analiz

İlk deneyde, seçilen zaman dilimleri $T_t = 1930 - 1939$ ve $T_q = 1980 - 1989$ için doğrusal dönüşüm matrisi **Q**'yu, CBOW ve SVD vektör yöntemleri ayrı ayrı kullanılarak elde edildi. Çizelge 3.1 ve Çizelge 3.2, sırasıyla CBOW ve SVD vektör algoritmaları için sonuçları göstermektedir. OP+SC yönteminin Recall@1, Recall@10 ve MRR metriklerinde diğer yöntemlere göre daha iyi performans sergilediği gözlemlendi. Recall@100 metriği açısından, OP yöntemi SVD vektörlerinde OP+SC yönteminden daha iyi performans gösterirken, CBOW vektörlerinde OP+SC yöntemiyle aynı skoru elde etmiştir. LT yöntemi ise tüm durumlarda en düşük skoru alarak önerilen yöntemlerin etkinliğini ortaya koymaktadır.

Çizelge 3.1 : 1930−1939 hedef dönemi ile 1980−1989 sorgu dönemi arasında farklı yöntemlerin CBOW vektörleri kullanıldığında performansları.

	Recall@1	Recall@10	Recall@100	MRR
LT	0,09	0,34	0,61	0,17
OP	0,33	0,71	**0,84**	0,45
OP+SC	**0,35**	**0,74**	**0,84**	**0,84**

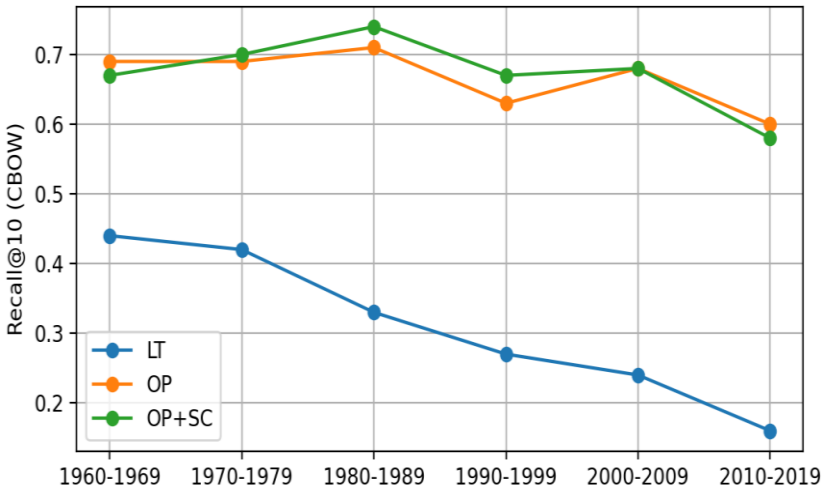
Çizelge 3.2 : 1930−1939 hedef dönemi ile 1980−1989 sorgu dönemi arasında farklı yöntemlerin SVD vektörleri kullanıldığında performansları.

	Recall@1	Recall@10	Recall@100	MRR
LT	0,23	0,59	0,82	0,34
OP	0,32	0,66	**0,86**	0,44
OP+SC	**0,33**	**0,70**	0,85	**0,85**

Kelime vektör algoritmalarının problem üzerindeki etkisiyle ilgili olarak, SVD vektörleri kullanıldığında OP ve OP+SC yöntemlerinin Recall@1 ve Recall@10 skorlarının azaldığı, ancak Recall@100 skorlarının arttığı gözlemlendi. Bununla birlikte. LT yönteminin performansı, SVD vektörleri kullanıldığında tüm metriklerde belirgin bir şekilde iyileşmiştir.

Bir sonraki deneyde, temel zaman diliminden hedef zaman dilimine olan zamansal mesafenin sonuçları nasıl etkilediği incelenmektedir. Bu bağlamda, hedef zaman

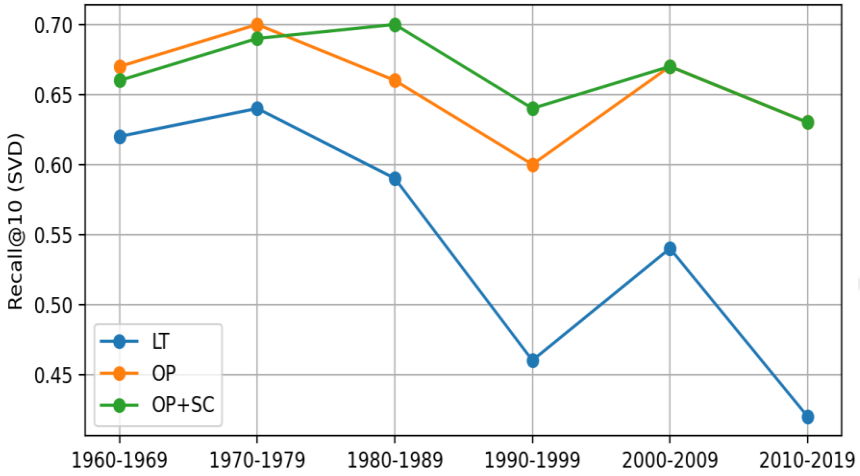
dilimi $T_B = 1930 - 1939$ ve sorgu zaman dilimleri ise sırasıyla $T_{q_1} = 1960 - 1969$, $T_{q_2} = 1970 - 1979$, $T_{q_3} = 1980 - 1989$, $T_{q_4} = 1990 - 1999$, $T_{q_5} = 2000 - 2009$ ve $T_{q_6} = 2010 - 2019$ olarak seçildi. Sorgu zaman dilimlerini belirlerken, karşılaştırmaların adil olması için hedef dilimler arasındaki token sayısının benzer olmasına dikkat edildi. Daha sonra, test kümesindeki (w,w') kelime ikililerini filtreleyerek ve bu sayede seçilen ikililerde w kelimesinin tüm hedef zaman dilimlerinde olmasını sağlayarak sonuçlar doğrudan karşılaştırılabilir hale getirildi. Sonuç olarak 221 kelime ikilisinde 220 kelime ikilisi bu şartı sağladı. LT, OP ve OP+SC yöntemlerinin farklı zaman dilimlerine karşı Recall@10 değerleri hesaplandı. Şekil 3.2 ve Şekil 3.3, sırasıyla CBOW ve SVD vektörleri için deney sonuçlarını göstermektedir.



Şekil 3.2 : CBOW vektörler ile farklı sorgu zaman dilimlerinde yöntemlerin Recall@10 değerleri.

İlk olarak, sonuçlar OP ve OP+SC yöntemlerinin LT'ye kıyasla zamansal farklılıklara karşı daha stabil olduğunu göstermektedir. İkinci olarak, önerdiğimiz yöntemlerle hem CBOW hem de SVD vektörlerinde benzer sonuçlar elde edilmiştir. LT'nin performansı hedef dönem ile sorgu dönem arasındaki zaman farkı arttıkça azalma eğilimi gösterirken, OP ve OP+SC yöntemlerinin 1960'lar ile 1980'ler arasındaki sorgu zaman dilimlerinde benzer kalmaktadır. Ancak, daha sonraki dönemlerde Recall@10 değerlerinde bir düşüş göstermektedir. Bu durum, dilin kendine özgü özelliklerinin bir sonucu olabilir. Özellikle, hedef dönem ile sorgu dönemi arasındaki mesafe arttıkça, iki dönem arasındaki kelime dağarcığının benzerliği azalmaktadır. Önceki zaman dilimlerinden bazı kelimeler, zaman geçtikçe

dilden yok olmaktadır. Sonuç olarak, iki dönem arasında paylaşılan kelimelerin sayısı azalmaktadır. Bu durum, dolayısıyla, Eşitlik (3.1) ve Eşitlik (3.2) üzerinden düşünüldüğünde, **Q** ve **M**'yi öğrenmek için kullanılan verinin miktarının azalmasına neden olmaktadır. Bu da doğrudan dönüşüm matrislerinin kalitesini düşürmektedir. Ayrıca, kelimeler zamanla semantik değişimlere uğramaktadır.



Şekil 3.3 : SVD vektörleri ile farklı sorgu zaman dilimlerinde yöntemlerin Recall@10 değerleri.

3.4 Sınırlamalar

Deneylerde, modern Türkçe kelimelerin eski karşılıklarını bulmak için önerilen yöntemlerin etkili olduğu gösterilmektedir. Ancak bu çalışma birkaç farklı açıdan geliştirilebilir. İlk olarak, Turkronicles veri seti, veri setinin oluşturulması sırasında belgelerden metinleri çıkarmak için kullanılan yöntemlerin sınırlamaları nedeniyle ortaya çıkan birçok gürültülü kelime içermektedir. Bu kelimeler, bir kelime için benzerlik kümesinin oluşturulmasını etkilemektedir. Örneğin, *vesik* kelimesinin, muhtemelen *vesika* kelimesinin yanlış yazılmış bir versiyonu olduğunu, OP yöntemi kullanıldığında *belge* kelimesi için en benzer 10 kelime arasında bulunmuştur. Bu nedenle, aslında doğru olan ancak bu gürültülü kelimeler nedeniyle yanlış kabul edilen birçok kelime olabilir. Ayrıca, gürültülü kelimeler, önerilen yöntemlerle bulunan dönüşüm matrislerinin kalitesini etkilemektedir.

Bir diğer eksiklik ise Spearman korelasyonunu hesaplamak için unigram frekansının kullanılıyor olmasıdır. Bu yöntem, kelimenin bağlam kelimeleriyle birlikte görülme sıklığını da dikkate alarak geliştirilebilir.

Ayrıca iki uzay arasında tek bir global dönüşüm matrisi bulmak kelimelerin kendi içindeki dinamikleri göz ardı edecek bir durumdur. Bu ilgili kelimeler için lokal dönüşüm matrisleri oluşturulabilinir.

Son olarak, deneyler tek bir diyakronik korpus kullanılarak oluşturulan bir test kümesine dayanmaktadır. *Turkronicles*, Türkçe için en büyük ve kapsamlı diyakronik korpus olsa da korpustaki belgeler Türkçe dilinin tüm özelliklerini yansıtmamakta ve yalnızca son 100 yılı kapsamaktadır. Bu nedenle, bu tür bir diyakronik korpus oluşturulduğunda, deneylerimizin daha geniş ve kapsamlı bir veri kümesinde tekrarlanması deneylerin daha sağlıklı sonuç vermesini sağlayacaktır. Ayrıca, manuel olarak oluşturulan test kümesinin genişletilmesi, daha güvenilir sonuçlar elde etmek açısından faydalı olacaktır.

4. TÜRKÇE'DE EŞ ANLAMLI KELİMELERİN ANLAM GENİŞLEMESİ VE DARALMASI

Doğal dillerin en önemli özelliklerinden birisi zamanla değişmesidir. Her dil kendini oluşturan bileşenler üzerinde farklı açılardan değişime uğrar. Bu değişimlerin birçok farklı sebebi vardır. Bunlar başlıca kültürel, politik, tarihsel ve dil bilimsel olarak kategorilere ayrılır. Dil içinde var olduğu toplumu doğrudan yansıttığı için toplum üzerindeki değişimler doğrudan dil üzerinde etki göstermeye başlar. Bu etki dilin semantik öğelerinde değişime yol açabilir. Bréal (1897) semantik kavramını ilk ortaya atmasından bu zamana kadar dil bilimciler semantik değişimin mekanizmalarını incelemişler sınıflandırmışlar. Bu mekanizmalardan başlıca öne çıkanlar Metaforik Değişim, Ad Aktarması Yoluyla Değişim, Genişleme, Daralma, Anlam İyileşmesi, ve Anlam Kötüleşmesi'dir (Bybee, 2015). Anlam değişimi dilin ve anlamın sahip olduğu bazı özellikler sayesinde meydana gelir. McMahon (1994) anlam değişimini mümkün kılan olan üç farklı koşulu belirtir. Bunların ilki kelimelerin çok anlamlı olma durumudur. Kelimeler insanlar tarafından farklı bağlamlarda kullanıldıkça yeni anlamlar kazanmaya başlar ve bu anlamlar birbiriyle genellikle ilişkilidir. İkinci koşul ise dilin aktarılma aşamasında kesikli bir şekilde alıcıya geçmesidir. Buna örnek olarak, çocuklar anne ve babalarından ve çevrelerinden dilin gramer kurallarını tam olarak kavrayamaz ve kendi kurallarını oluşturur. Bu da dilde anlamsal değişime sebep olur. Son olarak göstergenin (sign) rastgeleliği anlam değişiminin olması için gerekli bir koşuldur. Gösterge gösteren (signifier) ve gösterenden (signified) oluşur. Bu iki bileşen arasındaki bağ keyfidir. Zamanla ikisi birbirinden bağımsız olarak değişebilir.

Genişleme, bir kelimenin anlamının anlam uzayında daha büyük bir alana yayılması, anlam daralması ise bir kelimenin anlamının zaman geçtikçe anlam uzayında nispeten daha küçük bir bölgeye çekilmesidir. Örneğin Türkçe'de *oğlan* kelimesi eski kullanımlarda hem kız hem de erkek çocuk anlamına gelirken zamanla yalnızca erkek çocuk anlamını taşımaya başlamıştır (Bahattin, 2003). Daralma ve Genişleme'nin sebeplerinden bir tanesi, bir kelimenin başka bir kelimenin anlam

alanına girmesi olduğu düşünülmektedir (Bybee, 2015). Bu alana ait diğer kelimelerin artık anlamı eskisi kadar aktaramamasına sebep olur ve daha niş bir anlam alanında etkili olur. Böylelikle konseptlere karşılık gelen kategorileri tanımlayan kelime kümelerinde değişiklikler olur. Anlam alanlarına yeni kelimelerin girmesine neden olacak durumlara örnek olarak diller arası etkileşimler ve neolojizm gösterilebilir. (Bréal, 1897; Bybee, 2015; McMahon, 1994)

Türkçe anlam genişlemesini ve daralmasının kısa bir zaman içerisinde gözlemlenebileceği dinamiklere sahiptir. Türk Dil Devrimi'nin başlangıcından itibaren Türkçe'yi arileştirmek adına Arapça ve Farsça kökenli birçok kelime yerini Türkçe kökenli yeni kelimelere bıraktı. Türkçe'de bu sebeple birçok neolojizm örneği bulunur. Yukarıdaki tanımlar ve örnekler bu uğraşının semantik alanları değiştirmeye iteceğini gösterir. Çünkü böylelikle Arapça ve Farsça kelimelerin yerine eşanlamlı bir kelime koyarak aslında bu kelimelerin semantik alanlarına doğrudan bir müdahale gerçekleşir. Bu da bu alanlardaki kelimelerin anlam genişlemesi ve daralması değişikliklerine sebep olabilir. Buna uygun olarak araştırma sorusu şu şekilde oluşturuldu: Türkçe'de bir kelime başka bir kelimenin anlam alanına girdiğinde, bu anlam alanını temsil eden diğer bir kelimenin anlamı daralırken diğerinin genişliyor veya nispeten sabit kalıyor mu?

Bu araştırma sorusunu temel alınarak çalışmanın bu bölümünde eş anlamlı kelimelerin semantik uzayda anlam genişlemesi ve daralması açısından birbirlerine göre hareketlerini incelenmektedir. Bunun için ilk olarak eş veya yakın anlama gelen kelimelerin anlam genişlemesi ve daralması bir kelime ilişki çizgesi (word association graph) ile modellendi. Ardından anlam değişimini ölçmek için çizge teorisinde bir merkeziyet ölçütü olan Ara Merkezilik (Betweenness Centrality) (C_B) kullanıldı (Freeman, 1977). Ardından Güncel Türkçe Sözlük kullanılarak eş anlamlı kelimeler tespit edildi. Sonrasında kelime ikililerinin 1930 ve 2010 yılları arasındaki zaman dilimlerindeki C_B değerleri ölçüldü ve her bir kelime için zaman serisi oluşturuldu. Zaman serileri kullanılarak bu eşanlamlıların birbirini genişleme ve daralma açısından hangi yönde etkilediğini ölçmek için Sperman Korelasyon Katsayısı hesaplandı. Son olarak iki kelimenin birbirine etkisini sayısal olarak göstermek adına bir Doğrusal Karma Model (LMM) eğitildi. Sonuçlar analiz dildiğinde eş anlamlı kelimelerin daralma ve genişleme açısından birbirlerine göre zıt bir eğilimde olduğu istatistiksel olarak anlamlı bir şekilde gözlemlendi.

4.1 Problem Tanımı

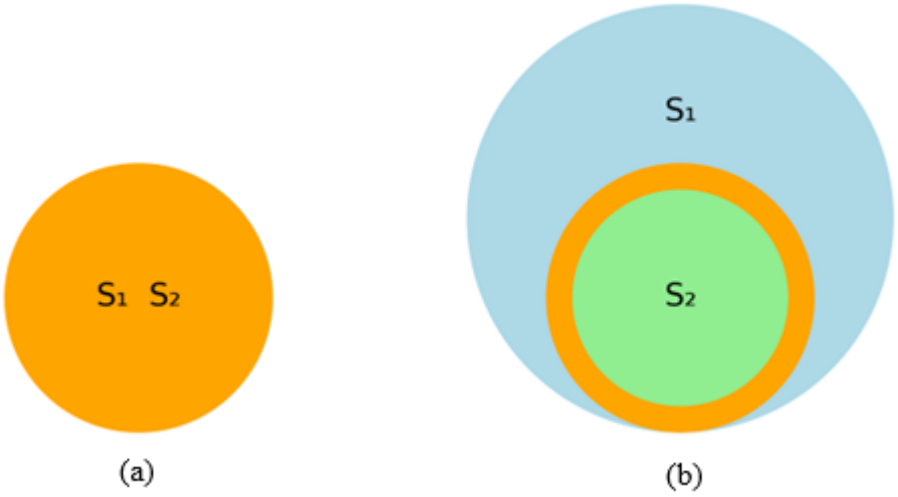
Bir kelimenin anlamını genişleterek semantik değişime uğramasına genişleme, anlam olarak daha dar bir alana denk gelmesine ise daralma denir. Eğer bir kelime başka bir kelimenin anlam alanına girerse, bu anlam alanındaki başka bir kelimenin anlamını daraltmasına sebep olabilir. Bir noktada aynı anlam alanındaki iki kelime eşanlamlı hale gelir. Anlamını genişleten kelime eski anlamıyla bir ilişkisi olacak şekilde yeni anlam kazanır. Anlamı daralan kelime ise yine ilişkili bir biçimde anlamını daraltır. Örneğin, İngilizce'de germen kökenli *hound* (tazı) kelimesi bir zamanlar evcilleştirilmiş köpek anlamına gelirken İskandinav kökenli *dog* (köpek) kelimesinin dile girmesiyle anlamı daralmış ve av köpeği ile ilgili anlam alanına çekilmiştir (Bybee, 2015).

Şekil 4.1'de anlam genişlemesi ve daralması şematik olarak gösterilmiştir. Burada sol tarafta ideal durumda aynı semantik alana düşen iki kelimenin, S_1 ve S_2, herhangi bir andaki durumları gösterilmektedir. Şekil 4.1 (a) incelendiğinde her iki kelime de başta aynı semantik alana karşılık gelir. Pratik olarak, bu iki kelime dil içinde kullanımlarında aynı anlamları taşıması beklenir. Fakat bu durum gerçeğe aykırıdır. Bir kelimenin diğeriyle tam olarak aynı anlama gelmesi oldukça nadir rastlanan bir durumdur. Bu sebeple figürler sadece genişleme ve daralmanın mantığının gösterilmesi amacıyla oluşturulmuştur. Şekil 4.1 (b)'de ise farklı bağlamlarda kullanılan S_1 'in anlam alanının genişlediği (açık mavi ile gösterilen alan) ve S_2'nin karşılık geldiği anlam alanının daraldığı (açık yeşil ile gösterilen alan) görülüyor. Burada dikkat edilmesi gereken nokta S_1 genişlerken eski anlamını da içermesi. Yani bir kelime anlamını genişletirken eski karşılık geldiği anlamı da nispeten karşılayabilmesi gerekir. Bu sebeple bir önceki anlamı ifade eden turuncu bölge açık mavi bölgenin içinde gösterilmiştir. Bunun yanında bir ihtimalde birinin anlam alanı göreceli sabit kalırken diğeri anlam alanını daraltmasıdır.

Buradaki amaç aynı anlam alanına düşen iki kelimenin zamanla anlam değişimi açısında nasıl bir hareket sergilediğini ölçmektir. Farklı bir deyişle ifade etmek gerekirse iki eş anlamlı kelimenin anlamının zaman sürecinde birbirlerine göre genişleme ve daralma açısından hangi yönde değiştiğini göstermektir.

4.2 Modelleme

Problemin tanımından yola çıkarak genişleme daralma dinamiklerini iyi şekilde yansıtacak yapı bir yönsüz kelime ilişki çizgesidir. Bu kelime ilişki çizgesinde her bir düğüm birer kelime kelimeler arasındaki ağırlıklı bağlantı ise kelimelerin arasındaki

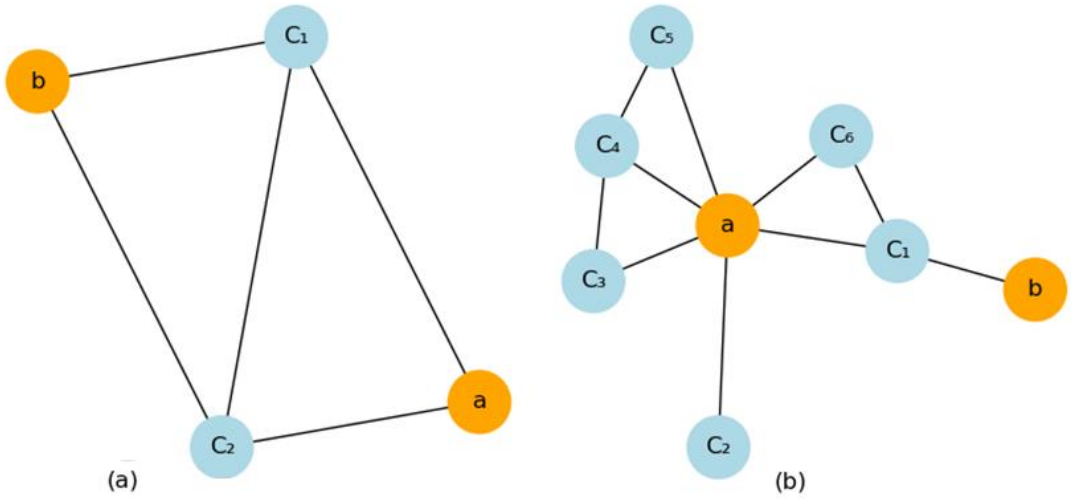


Şekil 4.1 : Aynı anlam alanına denk gelen kelimelerin zamanla değişiminin gösterimi.

ilişkinin "gücünü" ifade eder. Formal bir tanımla $G = (V, E)$ bir kelime ilişki çizgesi olsun. V bu çizgede kelimelere karşılık gelen düğümler ve E ise kelimeler arasındaki ilişkiyi temsil eden PPMI değerleriyle ağırlıklandırılan bağlantılardır. Genişleme ve daralma dinamiklerini modellemeden önce kelimenin anlam alanın çizgede nasıl bir yapıya karşılık geldiği tanımlanması gerekir. Anlam alanı bağlam ile ilişkili olduğu için dolaylı olarak genişleme ve daralma bağlam üzerinden gözlemlenebilinir. Bu bilgileri göz önünde bulundurursak kelimelerin metin içerisinde kullanıldığı bağlamlar, ilgili kelimenin çizgedeki doğrudan komşuluğuna denk gelir. Yani bir kelimenin anlam alanı, dolaylı olarak bağlamı, ona karşılık gelen düğümün komşularıyla birlikte oluşturduğu alt çizgedir. Daha formal bir biçimde ifade edecek olursak $u \in V$ kelimesinin anlam alanını oluşturan alt çizge $G_u = (V_u, E_u)$'dur. Burada $V_u = \{v : PPMI(u, v) > 0,\ \forall v \in V\} \cup u$ şeklindedir. Düğümler arasındaki bağlantılar ise $E_u = \{(x, y) : x, y \in V_u, (x, y) \in E\}$ elde edilir. u kelimesinin yanında bir de eşanlamlısı olan $v \in V$ kelimesi olsun. Bu kelimelerin çizgedeki ortak komşularının göreceli olarak fazla olmasını beklemek makuldür. Fakat kelimelerin

kullanım sıklığından ötürü farklılıklar ortaya çıkabilir. Bu sebeple u ve v eşanlamlı kelimelerinin karşılık geldiği bağlamları bulmak adına, bu kelimelerin komşuluklarıyla birlikte oluşturduğu alt çizgenin birleşimi alınır, $G_M = G_u \cup G_v = (V_M, E_M)$. G_M u ve v'nin karşılık geldiği bağlamları ifade eder. Bu birleştirme işlemiyle oluşan bağlam, çizgenin tümünde izole bir şekilde düşünülür. Bu izolasyon hesaplama aşamasında diğer semantik alanların bu iki kelime ve onların anlam alanı üzerindeki gürültülerini azaltır.

Problem tanımından yola çıkarak genişleme ve daralma dinamikleri incelendiğinde anlamı genişleyen bir kelimenin kullanıldığı bağlam sayısında artış gözlenmesi beklenir. Dolayısıyla bu kelimenin çizge üzerindeki bağlı olduğu kelime sayısı artacaktır. Anlamı daralan bir kelime için tersi durum geçerlidir. Bu durum çizge üzerinde Şekil 4.2'de gösterilmiştir.



Şekil 4.2 **:** Semantik ilişkileri temsil eden çizgenin zaman içindeki değişimi.

Şekil 4.2 (a)'da iki eşanlamlı kelime a, b başlangıçta aynı anlam alanına denk gelmektedir. Zaman geçtikçe a kelimesi daha farklı bağlamlarda kullanılarak anlamını genişletmiş, bu sırada b kelimesi daha az bağlamda kullanıldığı için anlamı daralmış şeklinde düşünülebilir. Böylelikle Şekil 4.2 (a)'daki çizge, Şekil 4.2 (b)'deki duruma gelmiştir. Bu gözlemden yola çıkarak anlamı genişleyen kelimenin aslında anlam uzayında daha merkezcil bir konuma geldiği görülür. Semantik yakınlık açısından bakıldığında anlamı genişleyen kelime farklı semantik alanlardaki kelimeleri dolaylı olarak birbirine yaklaştırarak aralarındaki semantik uzaklığını azaltır. Yani anlamı genişleyen kelimenin bağlamındaki kelimeler ile genişlemeden

önceki kelimeler daha ilişkili hale gelir. Buradaki doğal varsayım ise iki kelime arasındaki ilişkinin çizge üzerindeki mesafe arttıkça azalmasıdır. Bunun yanında anlamı genişleyen taraf hiyerarşik ve ideal olarak anlamı daralan kelimenin bağlamını kapsaması gerekir. Bu durumda anlamı daralan kelime izole anlam alanında daha düşük merkeziyete sahip olacaktır. Bu gözlemden yola çıkarak bir kelimenin merkeziyetini Ara Merkezilik (C_B) değeri ile bağdaştırdık. C_B ölçütü bir çizgede bulunan bir düğümün, tüm düğümler arasında en kısa yol hesaplandığında o düğümün üstünden geçen en kısa yolun sayısının tüm düğümler arasındaki en kısa yol sayısına oranıdır. Aşağıdaki eşitlikte gösterilmiştir.

$$C_B(v) = \sum_{s,t \in V_M} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{4.1}$$

Burada $C_B(v)$, $v \in V_M$ kelimesine karşılık gelen düğümün merkeziyet değerini temsil eder. σ_{st} ise s, t düğümleri arasındaki en kısa yolların sayısı ve $\sigma_{st}(v)$ v düğümü üzerinden geçen $s - t$ arasındaki en kısa yolların sayısıdır. Ardından bu ölçüt $1/(N-1)(N-2)$ değeri ile normalize edilir.

Zaman geçtikçe eşanlamlı kelime çiftindeki bir kelimenin C_B değerinin nispeten sabit veya artarken diğer kelimenin C_B değerinde azalma olması beklenir. Yani aslında iki kelimenin C_B değerlerini temsil eden zaman serileri arasında negatif veya 0 korelasyon mevcuttur. Bu korelasyonu ise Spearman korelasyon katsayısı ile ölçmek problemimiz için uygundur.

4.3 Önişleme

Turkronicles korpusundaki belgeler, orijinal halinde ya basılı ya da dijital olarak oluşturulmuştur. Basılı belgeler, orijinal materyalin taranmasıyla dijitalleştirilmiştir. Dijitalleştirme sürecinin önemli bir kısmı, metin verisi oluşturmak için Optik Karakter Tanıma (OCR) yazılımını kullanmaktır. Ancak, Bölüm 2'de de belirtildiği üzere, belgelerin nispeten eski ve fiziksel olarak hasarlı olması ve OCR yazılımlarının gürültüye karşı hassasiyeti nedeniyle, korpusta hatalı kelimeler ortaya çıkmaktadır.

OCR sonrası hataları düzeltme ile ilgili doğal dil işleme literatüründe birçok farklı çalışma bulunmaktadır. Bu yöntemler metodolojilerine göre kategorilere ayrıldığında

Kitle Kaynak Kullanımı (Crowd Sourcing), Sonlu Durumlu Dönüştürücü (Finite State Transducer), Diziden Diziye Modeller (Sequence-to-Sequence Models), ve LLM'leri kullanan çalışmalar şeklinde ana başlıklar ortaya çıkar (Nguyen ve diğ., 2021). Bu çalışmadaki deneyler yapılırken Mokhtar ve diğ. (2018) yaptığı çalışma temel alınıp hataları gidermek için bir Kodlayıcı-Çözücü (Encoder-Decoder) modeli eğitildi. Bu yaklaşımda OCR hatalarını düzeltme problemi bir MT problemi olarak düşünüldü. Yani, bu hataları kaynak metinden hedef metne dönüşüm olarak ele alarak, yanlış karakterlerin doğru karşılıklarıyla değiştirilmesini sağlayacak şekilde bir makine MT modeli geliştirildi. Referans metin, OCR çıktısı olan ve hatalı kelimeler içeren metin; hedef metin ise bu hatalı kelimelerin düzeltildiği metin olarak tanımlanır. Bu şekilde çalışan bir modelin eğitilmesi için paralel bir korpusa ihtiyaç duyulur. Fakat Türkçe'de literatür araştırmasının gösterdiği kadarıyla böyle bir paralel korpus bulunmamaktadır.

Tamamen insanlar tarafından etiketlenip oluşturulacak büyük bir paralel korpusun maliyeti oldukça yüksek olacaktır. Bu nedenle, daha az maliyetli ve daha hızlı bir çözüm sunmak için otomatik bir yöntem ile paralel korpus oluşturuldu. İlk olarak *Turkronicles*'da bulunan belgeler cümlelere bölütlendi. Cümleleri bölütlerken bir açık kaynak Python kütüphanesi olan *nltk*[16] kullanıldı. Ardından içinde hiçbir hatalı kelime olmayan cümleleri ve hatalı cümleleri Öztürel ve diğ. (2019) morfolojik analiz aracı kullanılarak tespit edildi. Sonrasında içinde hatalı kelime olan cümlelerden her 10 yıllık zaman diliminden olabildiğince eşit sayıda olacak şekilde cümleler seçildi. Toplam 309856 hatalı cümle *gpt4o* LLM'i kullanarak düzeltildi. Ardından cümle sayısını arttırmak için sentetik şekilde cümlelerde hatalar oluşturuldu. Bu işlemi yapmak için önceden tespit edilen ve hata içermeyen cümlelerin kelimeleri rastgele karakter ekleme, silme ve değiştirme işlemleri ile bozuldu. Böylelikle son durumda 535364 cümlelik paralel bir korpus oluşturuldu. Çizelge 4.1 otomatik olarak oluşturulan paralel korpustan bir kesit göstermektedir. Paralel korpus oluşturulduktan sonra %5 validasyon ve %95 eğitim kümesi olacak şekilde cümleler ayrıştırıldı. Modelin performansının değerlendirilmesi için, *Turkronicles* belgelerinden rastgele seçilmiş hatalı cümleleri içeren bir test kümesi hazırlandı. İlk olarak, yazım hatası içeren bir kelimeler belirlendi ve ardından bu

[16] https://www.nltk.org/

hatalı kelimeyi içeren en fazla 2 cümle test kümesine dahil edildi. Cümlelerdeki yazım hatalarını işaretlendi ve hatalı kelimeler belgelerin ilgili kısmı ile manuel biçimde karşılaştırılarak düzeltildi.

Çizelge 4.1 : OCR hataları için tam otomatik olarak üretilmiş paralel korpustan bir kesit.

Blı buçuk dönüm arazi bir köyIüye düşeıse bu köyde hayat vardır deneneız	Bir buçuk dönüm arazi bir köylüye düşerse bu köyde hayat vardır denemez.
Gemlike yapılacak demir yolu hatta Eskişehir içiii çok hayatidir	Gemlik'e yapılacak demir yolu hatta Eskişehir için çok hayatidir
Devletin büiün gelir ve şiderlerini, Anayasa gereği, Türkiye Büyük Millet Meclisi dcnetlemek durumdadır	Devletin bütün gelir ve giderlerini, Anayasa gereği, Türkiye Büyük Millet Meclisi denetlemek durumdadır.
Bcg sene bitii	Beş sene bitti
Sinzin pooülist politika üyguladığınızı ifade etmiyor	Sizin popülist politika uyguladığınızı ifade etmiyor

Test verisini hazırlama sürecinde veri seti ile ilgili şu bulguları gözlemledik: Bazı belgeler muhtemelen İngilizce için eğitilmiş OCR yazılımları kullanılarak taranmış. Bu nedenle, bazı Türkçe karakterler, özellikle *ü, ş, ç, ö*, görsel olarak benzer İngilizce harfler olarak OCR yazılımları tarafından yanlış yorumlanmış. Örneğin, belgelerdeki *ü* harfi *ii* veya *u* olarak algılanmış. Ayrıca, bazı belgeler ciddi şekilde hasar görmüş. Hasarlı bölümlerdeki yazılar kontrol aşamasında anlaşılması oldukça güç hale gelmiştir. Bunların dışında, belgelerin yazarlarından kaynaklanan yazım hataları da cümlelerde mevcuttur.

Çizelge 4.2 test verisinin genel istatistiklerini göstermektedir. Test kümesi 999 hatalı cümle içermektedir. Toplamda 11019 kelime bulunmaktadır. Bunların 2760 tanesi yazım hatası içeriyor ve bu da yaklaşık %25'e tekabül etmektedir. Test verisinin OCR hata miktarı iki farklı metrikte ölçüldü. Kelime Hata Oranı (WER) ve Karakter Hata Oranı (CER), ASR ve OCR sistemlerinin performansını değerlendirmek için kullanılan ölçütlerdir. WER metriği $\frac{Y+S+E}{N}$ formülü ile hesaplanır. Burada Y referans cümledeki bir kelimenin yerine gelen yanlış kelimelerin sayısı, S referans cümledeki bir kelimenin silinme sayısı, E referans cümleye yeni bir kelimenin eklenme sayısı ve N ise toplam kelime sayısıdır. CER metriği ise $\frac{YK+SK+EK}{N}$ şeklinde hesaplanır. YK referans cümledeki bir karakterin yerine geçen yanlış karakterlerin sayısı, SK referans cümledeki bir karakterin silinme sayısı, E referans

cümleye yeni bir karakterin eklenme sayısı ve N ise toplam karakter sayısıdır. Çizelge 4.2'de test verisinin WER değeri 0,250 ve CER değeri ise 0.056'dır. Ek olarak hatalı bir cümle ile düzeltilmiş versiyonu arasındaki Levenshtein Mesafesi hesaplanmıştır. Hem ekleme hem de silme masrafı 1 olarak belirlenirken, yerine koyma masrafı 2 olarak ayarlandı. Bu parametrelerle yapılan hesaplama neticesinde, cümle başına ortalama 7.75 düzenleme mesafesi olduğu gözlemlenmiştir.

Çizelge 4.2: Test verisi istatistiği

Cümle Sayısı	999
Kelime sayısı	11019
Hatalı Kelime Sayısı	2760
Kelime Hata Oranı (WER)	0,250
Karakter Hata Oranı (CER)	0,056
Ortalama Levenshtein Mesafesi	7,75

Literatür araştırmasına göre oluşturduğumuz test kümesinin tarihî Türkçe belgelerden oluşturulmuş ilk test kümesi olma özelliğini taşır. Bu test kümesi, Dil Modellerine (LM) dayanan OCR sonrası düzeltme gerçekleştiren sistemlerin performansının değerlendirilmesi aşamasında katkı sağlayacaktır. Ayrıca test kümesini açık erişim hale getirerek, OCR sonrası düzeltme sistemlerinin gelişimine ve değerlendirilmesine katkıda bulunmayı amaçlıyoruz.

Eğitim aşamasına geçilmeden önce, tüm cümleler *SentencePiece* algoritmasıyla tokenize edildi. Kelime dağarcığı boyutu 30000 olarak belirlendi.

OCR hatalarını düzeltmek amacıyla geliştirilmek istenen model, Kodlayıcı-Çözücü mimarisine dayanmaktadır. Kodlayıcı ve Çözücü bileşenlerinin her biri 6 adet Dönüştürücü (Transformer) bloğu içermektedir. Her Dönüştürücü bloğunda 4 adet Attention Head[17] mekanizması bulunmaktadır. Kelimelerin vektör boyutları $d = 512$ olarak belirlenmiştir. Transformer bloklarındaki ileri beslemeli ağların ara katmanlarının (hidden layer) boyutları ise $d_{ff} = 2048$ olarak tanımlanmıştır. Model hem anlam bütünlüğünü koruyacak şekilde cümleleri işlemekte hem de yanlış kelimeleri bağlamı dikkate alarak düzeltmeye odaklanmaktadır.

Model eğitildikten sonra, oluşturulan test kümesi ile modelin başarısı ölçüldü. Çizelge 4.3 test sonuçlarını göstermektedir. Çizelge 4.3, modelin düzeltme performansını WER ve CER metriklerinde değerlerini. WER (Kelime Hata Oranı),

[17] Dönüştürücü bloklarında kullanılan temel bileşenlerden biri.

orijinal test verisinde 0,250 iken düzeltme sonrasında 0,120'ye düşmüş ve bu durum %51,9 oranında bir iyileştirme sağlamıştır. CER ise başlangıçta 0,056 olarak ölçülmüş ve düzeltme sonrası 0,031'e düşerek %44,6 oranında bir iyileşme göstermiştir.

Çizelge 4.3 : Modelin test sonuçları WER ve CER metrikleriyle birlikte test verisi iyileştirme oranı.

	Orijinal Test Verisi	Düzeltme Sonuçları	İyileştirme Oranı
WER	0,250	0,120	%51,9
CER	0,056	0,031	%44,6

Elde edilen bulgulara dayanarak, modelin hem kelime hem de karakter seviyesindeki hata oranlarını önemli ölçüde azalttığını ve düzeltme performansının etkili olduğunu ortaya koyar.

Hata düzeltme işlemleri tamamlandıktan sonra ilk olarak cümleler tokenize edildi. Ardından tüm kelimeler küçük harfe dönüştürüldü. İçerisinde harf dışında herhangi bir karakter olan kelimeler kaldırıldı. Bölüm 1'de kullanılan morfolojik analiz aracı kullanılarak kelimeler sözlük formuna indirgendi. Eğer morfolojik analiz aracı başarısız olursa ilgili kelime analize ve modellere dahil edilmedi.

4.4 Veri Kümesi

Anlam genişlemesi ve daralmasını değerlendirmek için Güncel Türkçe Sözlük kullanılarak bir test kümesi oluşturuldu. İlk olarak sözlükte bulunan eş anlamlı kelimeler tespit edildi. Ardından bu kelimelerin bütün zaman dilimlerinde en az 50 kere geçtiğinden emin olacak şekilde eleme işlemi gerçekleştirildi. Frekans eşik değerinin sebebi ise bu kelimelerin bağlamlarının yeteri kadar kapsayıcı olmasını sağlamaktır. Böylelikle 673 adet eş veya yakın anlamlı kelime çifti elde edildi.

4.5 Deneyler

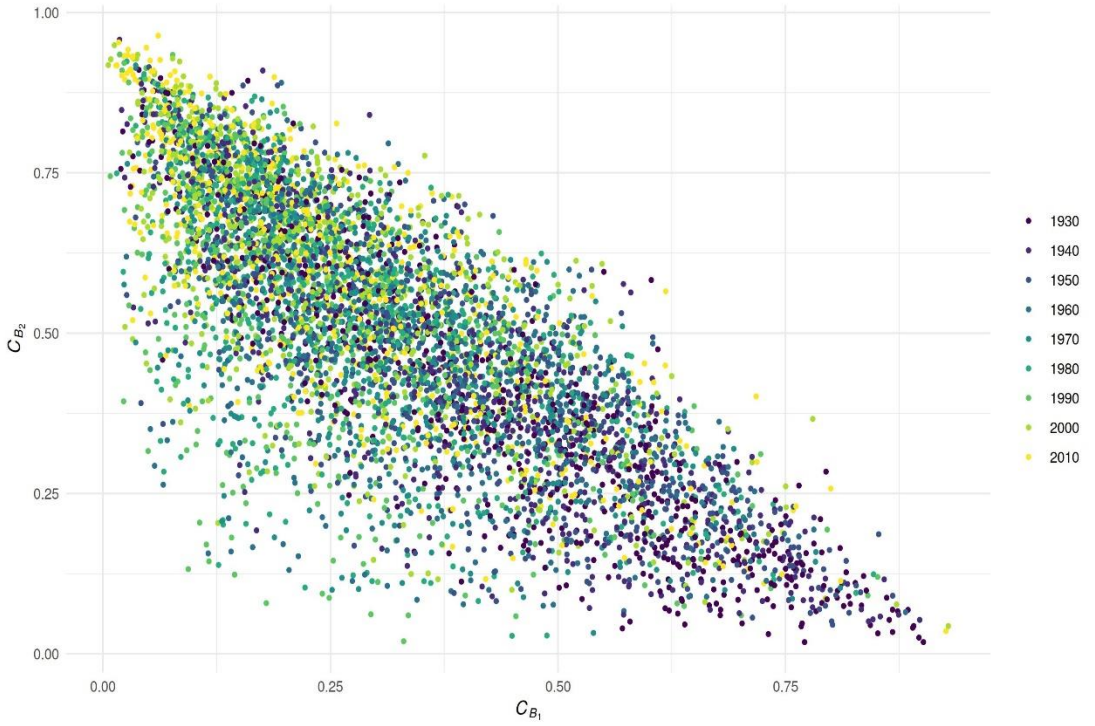
Önişleme aşamasından geçirilmiş *Turkronicles* korpusu ilk olarak 10 senelik zaman dilimlerine ayrıldı. 1920-1929 arası dokümanlar Osmanlı alfabesinde yazıldığı için yakın zamanda Latin Alfabesine çevrilmiştir ve token sayısı diğer zaman periyotlarına göre oldukça düşüktür. Ayrıca 2020-2022 zaman aralığı tamamlanmadığı için bu aralıkta da token sayısı diğerlerine nazaran oldukça azdır.

Bu sebeplerle 1920-1929 ve 2020-2022 zaman aralıklarında bulunan dokümanlar analize dahil edilmedi. Bu durumda analiz yapılan zaman aralığı 1930-2019 yılları arasını kapsamaktadır. Zaman aralığını belirledikten sonra her 10 zaman periyodu için kelime ilişki çizgesi oluşturuldu ve PPMI değerleri kullanılarak düğümler arası bağlantılar ağırlıklandırıldı. PPMI değerleri bulunurken bağlam penceresinin büyüklüğü sağda ve solda 2 token olmak üzere toplam 5 olarak seçildi. Ardından PPMI değeri 3'ten düşük olan bağlantılar bu çizgelerden kaldırıldı. Bu işlemi yaparak, yalnızca güçlü ve anlamlı bağlantıların kalması sağlandı. Böylece, modelin daha doğru ve tutarlı sonuçlar üretmesi hedeflendi. Bu işlemlerin sonunda her bir zaman dilimi için birer çizge elde edildi.

Çizgeler oluşturulduktan sonra test kümesindeki her bir eşanlamlı kelime çifti için, her bir zaman dilimine ait G_M alt çizgesini oluşturuldu. G_M çizgesi kullanılarak bir eşanlamlı çiftindeki her bir kelime için birer zaman serisi oluşturuldu. Bu zaman serilerinin elemanları, bir kelimenin ilgili zaman aralığındaki G_M çizgesi yardımıyla hesaplanan C_B değerine karşılık gelmektedir. Şekil 4.3, örneklerdeki eş anlamlı kelime çiftlerine ait kelimeleri Ara Merkezilik değerleri arasındaki ilişkiyi göstermektedir. C_{B_1}, bir eş anlamlı kelime çiftindeki ilk kelimenin Ara Merkezilik değeri, C_{B_2} ise ikinci kelimenin Ara Merkezilik değeridir. Eş anlamlı çiftlerini temsil eden noktalar alındığı zaman dilimine göre renklendirilmiştir. Grafik görsel olarak incelendiğinde ilk göze çarpan kısmın iki kelimenin birbirini negatif olarak etkileyecek şekilde doğrusal bir trend oluşturduğudur. Farklı zaman dilimlerini temsil eden renkler, bu negatif ilişkinin dönemler boyunca tutarlı olduğunu da ortaya koymaktadır. Ayrıca grafiğin iki ucuna bakıldığında aşağı uçta 1930'lardan noktaların daha yoğun olduğu, yukarı uçta ise 2010'lardan noktaların daha yoğunlukta olduğu gözlemlenmektedir. Bu da zaman diliminin eş anlamlı kelime grupları arasındaki etkisini göz önüne serer. Test kümesi içerisinde geçen her bir kelimenin zaman serileri oluşturulduktan sonra o kelimenin eş anlamlısının zaman serisiyle arasındaki Spearman Korelasyon Katsayısı ölçüldü. Negatif korelasyon, uygulanan modele göre bir kelimenin anlamı genişledikçe diğerinin daraldığını gösterir. Diğer taraftan, pozitif korelasyon, bir kelimenin anlamı genişledikçe diğerinin de benzer bir hareket ettiğine işaret eder.

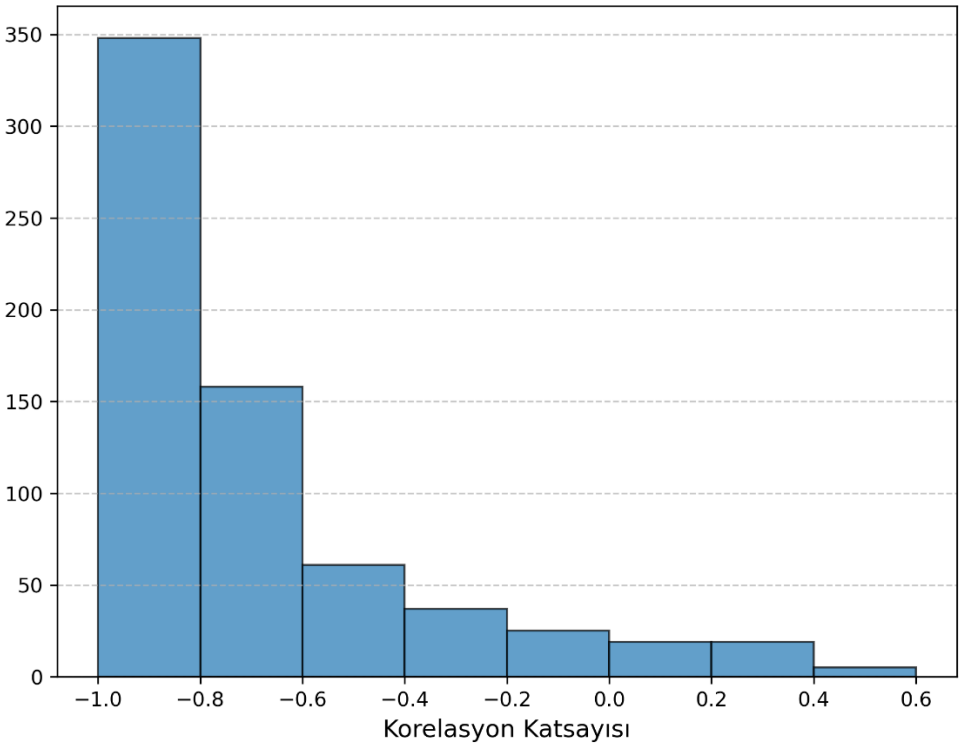
Korelasyon katsayılarının dağılımı Şekil 4.4'teki histogram grafiğinde gösterilmiştir.

İlk olarak korelasyon aralıkları $[-1.0, 0.8)$ çok güçlü negatif, $[-0.8, -0.6)$ güçlü



Şekil 4.3 **:** Aynı anlam alanındaki iki kelimenin birbirlerine göre $\boldsymbol{C_B}$ değerini gösteren grafik..

negatif, $[-0.6, -0.4)$, orta derecereli negatif, $[-0.4, -0.2)$ zayıf negatif, $[-0.2, 0.0)$ çok zayıf negatif, $[0.0, 0.2)$ çok zayıf pozitif, $[0.2, 0.4)$ zayıf pozitif, $[0.4, 0.6)$ orta dereceli pozitif, $[0.6, 0.8)$ güçlü pozitif, $[0.8, 1.0]$ çok güçlü pozitif olarak kategorilere ayrıldı. 673 eş anlamlı kelime çiftinin 630'unda negatif korelasyon gözlemlendi. Bunlardan 484'ü istatistiksel olarak anlamlıdır ($p < 0{,}05$). 348 kelime çiftinin çok güçlü negatif korelasyon grubunda olduğu görülüyor. Örneklerin çoğu çok güçlü ve güçlü negatif korelasyon grubundadır (%75). Ayrıca Bununla beraber korelasyon katsayılarının %93'ü negatif korelasyonludur. 673 örneğin 43'ü pozitif korelasyona sahiptir. Pozitif korelasyona sahip örnekler çok zayıf, zayıf, ve orta derece pozitif kategorilere dağılmıştır. Bu kelimeler arasında *hariç-dış*, *mis-güzel*, *mutlaka-kesin*, *nadir-seyrek*, *salim-esen*, *reşit-ergin* gibi sıfat ve zarf niteliğinde olan ikililerin var olduğunu görüyoruz. Özellikle sıfat ve zarfların bağlamsal esneklikleri ve çeşitlilikleri nedeniyle bu kelime ikilileri pozitif korelasyona sahip olmuş olabilir. Ayrıca pozitif korelasyon sahip eş anlamlı çiftlerinin hiç biri istatistiksel olarak anlamlı sonuç vermemiştir $(p > 0{,}05)$.



Şekil 4.4 : Eşanlamlı kelime çiftlerini merkeziyet değerlerinin korelasyon katsayısı histogram grafiği.

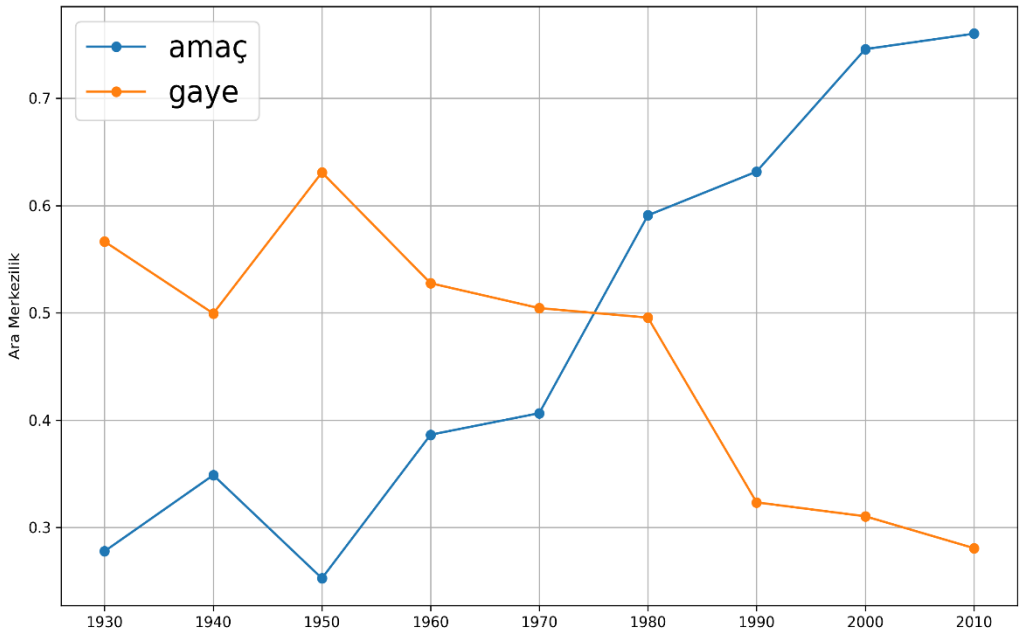
Bir kelimenin anlam genişlemesi ve daralmasında, eş anlamlı karşılığının anlam alanını ne kadar etkilediğini ölçmek için LMM kullanılarak bir analiz gerçekleştirildi. İlk olarak her bir eş anlamlı kelime çifti eşsiz bir şekilde tanımlanması için indekslendi. Her bir kelime çiftine ait 1930-2010 yıllarını kapsayacak şekilde merkeziyet değerlerinin ölçümü olan zaman serileri oluşturuldu. Bu sebeple global trendin yanında grup içi varyasyonları ölçmek adına modele her bir kelime çiftine karşılık gelecek şekilde rastgele etki katsayısı u_i dahil edildi. Ayrıca zaman periyotlarının genişleme/daralma üzerindeki etkisini ölçmek için, zamanı kategorik bir değişken olarak düşünülüp modele sabit etki β_t dahil edildi. Son durumda bu model Eşitlik 4.2'de gösterilmiştir.

$$C_{B_{1,i}}^{(t)} = \beta_t + \beta \cdot C_{B_{2,i}}^{(t)} + u_i + \epsilon_i^{(t)} \tag{4.2}$$

$C_{B_{1,i}}^{(t)}$, t zaman diliminde i'nci eşanlamlı çiftindeki ilk kelimenin C_B değeridir. $C_{B_{2,i}}^{(t)}$ ise t zaman diliminde i'nci eşanlamlı çiftindeki ikinci kelimenin C_B değeridir. $u_i \sim \mathcal{N}(0, \sigma_i^2)$ i'nci eşanlamlı çiftinin rastgele etki katsayısıdır. Bu modeli eğittiğimizde sonuçlar eş anlamlı kelime çiftlerinde semantik genişleme ve daralmanın dinamiklerini güçlü bir şekilde desteklemektedir. $C_{B_{2,i}}^{(t)}$'nin sabit etki katsayısı $\beta =$

$-0{,}692$ olarak bulunmuştur ($p < 0{,}001$). Bunun yanında, modelin R^2 değeri 0,90 olarak hesaplanmıştır. Bu değer modelin değişimdeki varyasyonun %90'ını açıklayabildiğini göstermektedir. Bu da kullanılan modelin güvenilirliğini ve uygunluğunu gösterir. Tüm bu veriler iki eşanlamlı kelime arasındaki daralma ve genişleme şeklindeki anlamsal değişimin güçlü bir negatif ilişki içinde olduğunun güçlü bir kanıtını sunar.

4.6 Örnek İncelemeler: *amaç-gaye, kuşak-nesil*



Şekil 4.5 : *amaç* ve *gaye* kelimelerinin yıllara göre Ara Merkezilik (C_B) değeri.

Modellemenin doğruluğunu incelemek için bilinen anlamı genişlemiş bazı kelimeler üzerinden incelemeler gerçekleştirdik. *amaç-gaye, alçak-kısa, kuşak-nesil* kelimeleri Türkçe Tarama Sözlüğü ve Güncel Türkçe Sözlük karşılaştırılarak anlamları genişlediği tespit edilmiştir (Özavşar, 2013). Bu kelime çiftlerinin ara merkezilik zaman serileri üzerinde yapılan analiz, eşanlamlı kelimelerden birinin anlam alanını genişletmesi durumunda diğerinin daraldığına işaret edecek şekilde dinamik etkileşimleri ortaya koymaktadır.

amaç kelimesi 18. Yüzyılda hedef, nişangâh[18] anlamına gelirken günümüz Türkçe'sinde anlamını genişleterek *gaye* anlamını taşımaya başlamıştır. Şekil 4.5 *amaç* ve kelimelerinin yıllara göre C_B değerleri görülmektedir. İlk yıllarda *gaye* daha yüksek ara merkezilik değeriyle geniş bir anlam alanına sahipken, 1960'tan sonra *amaç* kelimesinin bağlam çeşitliliği giderek artarak onu daha merkezî bir hale getirmiştir. Bu kelimelerin C_B zaman serileri arasında güçlü negatif korelasyon gözlemlenmiştir ($r = -0.95,\ p < 0.001$).

Benzer şekilde, *kuşak-nesil* eş anlamlı çifti için elde edilen güçlü negatif korelasyon ($r = -0.87,\ p = 0.0025$), birbirine zıt hareketi destekleyen bir başka örnektir. Başlangıçta *kuşak* daha yüksek C_B değerine sahipken, 1940'tan sonra *nesil* kelimesi giderek merkeziyet değerini arttırmış ve 1990'da *kuşak*'ı geçmiştir. Bu sırada *kuşak* kelimesi zaman boyunca anlamını daraltması C_B değerlerine yansımıştır.

4.7 Sınırlamalar

Anlam genişlemesi ve anlam daralmasını modellemek için PPMI çizgelerinden yararlanılmıştır. Modelleme aşamasında anlam alanı oluşturulurken iki eş anlamlı kelimenin komşularının oluşturduğu alt çizgelerin birleşimi alınmaktadır. Bu durum kelimelerin çok anlamlılık özelliklerini göz ardı edip bazı durumlarda uzak anlamların hesaba katılmasına sebep oluyor olabilir. Örneğin, eş sesli bir kelime (örneğin "yüz") farklı bağlamlarda kullanıldığından, izole anlam alanına uzak veya ilgisiz kelimeler girebilir. Bu sebeple iki kelimenin sadece ilgili anlam alanlarını daha iyi çıkartacak bir yöntem uygulanması modelin başarısını arttıracaktır.

Genişleme ve daralma dinamiğini ölçmek için C_B ölçütü kullanılmıştır. C_B hesaplanırken PPMI değerlerinden, alakasız kelimelerin elenmesi aşamasında yararlanılmıştır. Fakat bu ölçüt hesaplanırken herhangi bir şekilde bağlantıların ağırlıkları dikkate alınmıyor. Bu ağırlıklar kelimelerin bir anlamda birbirine bağlılıklarını ve güçlerini belirttiği için PPMI değerleri mesafe hesabına katıldığında model daha farklı sonuçlar verebilir.

Bağlam çeşitliliğini yansıtmak adına, her bir kelime çifti için belirli bir zaman diliminde bulunma şartı ve frekans için eşik değerler belirlenmiştir. Ancak,

[18] TDK Tarama Sözlüğü.

kelimelerin kendi bağlam çeşitliliği veride yeterince iyi yansıtılmamış olabilir. Bu durum, modelin sonuçlarını sınırlayabilecek bir faktördür. Bununla birlikte, bu çalışmanın tekrarlanması ve daha büyük, daha çeşitli bir veri kümesiyle uygulanması, elde edilen sonuçların doğruluğunu artırma potansiyeline sahiptir.

Analizler iki kelime üzerinden yapılmıştır. Bu durum sonuçları etkiliyor olabilir. Örneğin semantik alana üçüncü bir kelime girdiğinde, yeni giren kelime, diğerlerini daralmaya itebilir. Bu da sonuçlarda pozitif korelasyon şeklinde görülebilir.

5. SONUÇ VE ÖNERİLER

Türkçe, özellikle alfabenin değiştirilmesi ve ödünç kelimelerin Türkçe kökenli kelimelerle değiştirilmesi gibi girişimlerle son yüzyılda önemli değişikliklere maruz kalmıştır. Türkçe'nin bu dönüşümünü inceleyecek ve gelecekteki çalışmaları ve tarihi belgeleri işlemek için NLP araçlarının geliştirilmesini desteklemek amacıyla 1920'den 2022'ye kadar Resmî Gazete ve Türkiye Büyük Millet Meclisi kayıtlarından çıkarılan diyakronik bir korpus olan *Turkronicles*'ı oluşturuldu. Ayrıca, bu önemli konu üzerindeki çalışmaları desteklemek için farklı türlerde diyakronik hizalanmış vektörler, PPMI ve frekans matrisleri, dijitalleştirilmiş karşılıklar sözlüğü ve bir Python kütüphanesi kullanıma sunuldu. *Turkronicle*s ile erişime açılan diyakronik kaynakların, dildeki değişim üzerine odaklanan çalışmalar için yararlı bir kaynak olacağını düşünülmektedir. Türkçe için bir diyakronik korpusun oluşturulması, dilbilimsel araştırmalarda önemli bir boşluğu doldururken, gelecekte korpusumuzu gazeteler, edebi eserler ve kamu yayınları gibi diğer kaynaklardan metinler eklenerek genişletilmesi planlanmaktadır. Çeşitli kaynaklardan daha büyük bir diyakronik korpus oluşturduktan sonra, Tür Dil Devrimi'nin Türkçe üzerindeki etkisini inceleyen analizlerin genişletilmesi ve farklı veri kaynaklarını karşılaştırılması hedeflenmektedir.

Diyakronik korpus oluşturulduktan sonra modern Türkçe kelimelerin eski karşılıklarını bulmak için ortogonal dönüşümler temelinde iki yöntem önerildi. Yöntemler farklı vektör algoritmaları üzerinden karşılaştırılarak zaman dilimleri arasındaki farkların performansa etkisini değerlendirmek amacıyla çeşitli deneyler gerçekleştirildi. Yapılan analiz tarihsel süreçte modern Türkçe kelimelerin eski karşılıklarının bulunmasında ortogonal dönüşümün doğrusal dönüşümden daha iyi sonuç verdiğini göstermektedir. Ayrıca, sorgu kelimesinin ilgili hedef zaman dilimindeki benzerlik kümesini Spearman sıra korelasyon değerlerine göre sıralamanın, OP yönteminin performansını hem SVD vektörlerinde hem de CBOW vektörlerinde çoğu durumda artırdığı gözlemlendi.

Son olarak bir kelimenin anlamını genişleterek başka bir anlam alanına girdiğinde o anlam alanındaki başka bir kelimenin anlamının daralmasına sebep olup olmadığı incelendi. Bu değişimi en iyi şekilde gözlemlemek için analizler eşanlamlı kelimeler üzerinden gerçekleştirildi. İlk olarak problem PPMI çizgesi ile modellendi ve anlam değişikliği dinamiğini ölçmek için Ara Merkezilik ölçütü kullanıldı. Yaptığımız deneyler sonucunda eşanlamlı kelimelerin birbirlerine göre aslında zıt şekilde anlam değişikliğine uğradıklarını gösteren bulgular elde edildi. Ardından zaman serileri için bir LMM modeli eğitildi ve önceki deneylerde elde edilen bulgular pekiştirildi. Burada bulunan sonuçlar da eşanlamlı kelimelerin birisinin anlamı genişlerken diğerinin daralmasını destekler nitelikteydi.

Bu çalışmada Türkçe'nin zaman boyunca gelişimini anlamak ve doğal dil işleme araçlarıyla bu dönüşümü modellenmesi açısından önemli bulgular içermektedir. *Turkronicles* gibi diyakronik bir korpusun oluşturulması, sadece Türkçe'deki değişimlerin izlenmesine değil, aynı zamanda dilin değişimiyle ilgili daha geniş kapsamlı çalışmaların yapılmasına olanak sağlayacak bir altyapı sunmaktadır. Bu bağlamda düşünüldüğünde eşanlamlı kelimelerin diyakronik tespit edilmesi ve anlam genişlemesi ve daralması gibi dilin değişim mekanizmalarını modellemek için sunulan kaynakların, dil değişikliği üzerine yapılacak gelecek çalışmalar için bir temel oluşturacağını düşünüyoruz. Elde edilen sonuçlar hem metodolojik hem de dilbilimsel açıdan önemli katkılar sunarken, Türkçe ve diğer dillerdeki tarihsel dil değişimini anlamaya yönelik çalışmalara da ilham verecektir.

KAYNAKLAR

Aitchison, J. (2001). *Language Change: Progress Or Decay?* Cambridge University Press. https://books.google.com.tr/books?id=KjcYC0qp2PgC adresinden alındı

Aitchison, J. (2005). Language change. *The Routledge Companion to Semiotics and Linguistics* (s. 111–120). içinde Routledge.

Aksan, D. (1965). Türk Anlam Bilimine Giriş–Anlam Değişmleri. *Türk Dili Araştırmaları Yıllığı-Belleten, 13*, 167–184.

Aksan, Y., Aksan, M., Koltuksuz, A., Sezer, T., Mersinli, Ü., Demirhan, U. U., . . . others. (2012). Construction of the Turkish National Corpus (TNC). *LREC*, (s. 3223–3227).

Altinok, D. (2023). A Diverse Set of Freely Available Linguistic Resources for Turkish. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (s. 13739–13750).

Altintas, K., Can, F., & Patton, J. M. (2007). Language change quantification using time-separated parallel translations. *Literary and Linguistic Computing*, *22*(4), 375-393.

Altinok, D. (2023). A Diverse Set of Freely Available Linguistic Resources for Turkish. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (s. 13739–13750).

Aronoff, M., & Lindsay, M. (2016). Competition and the lexicon. *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della società di linguistica Italiana. Roma: Bulzoni Editore*, (s. 39–52).

Bahattin, S. A. (2003). Anlam değişmeleri üzerine artzamanlı bir inceleme. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 23*.

Bamman, D., & Crane, G. (2011). Measuring historical word sense variation. *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, (s. 1–10).

Bréal, M. (1897). *Essai de sémantique: (science des significations).* Hachette. https://books.google.com.tr/books?id=fWN0tQEACAAJ adresinden alındı

Bybee, J. (2015). *Language change.* Cambridge University Press.

Can, F., Kocberber, S., Balcik, E., Kaynak, C., Ocalan, H. C., & Vursavas, O. M. (2006). First large-scale information retrieval experiments on Turkish texts. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, (s. 627–628).

Can, F., & Patton, J. M. (2010). Change of word characteristics in 20th-Century Turkish literature: A statistical analysis. *Journal of quantitative linguistics*, *17*(3), 167-190.

Cassotti, P., De Pascale, S., & Tahmasebi, N. (2024). Using Synchronic Definitions and Semantic Relations to Classify Semantic Change Types. *arXiv preprint arXiv:2406.03452*.

Clauson, G., CLAUSON, G., & ÖZALAN, U. (2007). Türkçede sekizinci yüzyıldan önce kullanılan ekler. *Dil Araştırmaları, 1*, 185–196.

Closs Traugott, E. (1985). On regularity in semantic change.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora, 7*, 121–157.

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019). Training temporal word embeddings with a compass. *Proceedings of the AAAI conference on artificial intelligence*, *33*, s. 6326–6334.

Fischer, R. (1998). Lexical change in present-day English: A corpus-based study of the motivation, institutionalization, and productivity of creative neologisms.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry.*

Goel, A., & Kumaraguru, P. (2021). Detecting Lexical Semantic Change across Corpora with Smooth Manifolds (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*, s. 15783–15784.

Goodman, J. T. (2001). A bit of progress in language modeling. *Computer Speech & Language, 15*, 403-434. doi:10.1006/csla.2001.0174

Gökçe, H. (2010). Başkurt Türkçesinde gramatikalleşme örnekleri üzerine. *Türkoloji dergisi, 17*, 83–104.

Göksel, A., & Kerslake, C. (2004). *Turkish: A comprehensive grammar.* Routledge.

Güngör, O., Tiftikci, M., & Sönmez, Ç. (2018). A Corpus of Grand National Assembly of Turkish Parliament's Transcripts. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation.*

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Jespersen, O. (1922). *Language: Its Nature, Development, and Origin.* Allen and Unwin.

Johanson, L. (1998). Turkic languages. *Encyclopaedia Britannica.* içinde

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Second b.). Pearson Prentice Hall.

Jurgens, D., & Stevens, K. (2009). Event detection in blogs using temporal random indexing. *Proceedings of the Workshop on Events in Emerging Text Types*, (s. 9–16).

Kernighan, M. D., Church, K., & Gale, W. A. (1990). A spelling correction program based on a noisy channel model. *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics.*

Kolak, O., & Resnik, P. (2002). OCR error correction using a noisy channel model. *Proceedings of the second international conference on Human Language Technology Research*, (s. 257–262).

Kornfilt, J. (1997). *Turkish.* London & New York: Routledge.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th international conference on world wide web*, (s. 625–635).

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Lees, R. B. (1961). The phonology of modern standard Turkish. *(No Title)*.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics, 3*, 211–225.

Lewis, G. (1999). *The Turkish language reform: A catastrophic success: A catastrophic success.* OUP Oxford.

Lewis, G. (2000). *Turkish grammar*. Oxford University Press.

Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature, 449*, 713–716.

Llobet, R., Cerdan-Navarro, J.-R., Perez-Cortes, J.-C., & Arlandis, J. (2010). OCR post-processing using weighted finite-state transducers. *2010 20th International Conference on Pattern Recognition*, (s. 2021–2024).

Löfgren, V., & Dannélls, D. (2024). Post-OCR Correction of Digitized Swedish Newspapers with ByT5. *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, (s. 237–242).

Luu, K., Khashabi, D., Gururangan, S., Mandyam, K., & Smith, N. A. (2022). Time Waits for No One! Analysis and Challenges of Temporal Misalignment. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (s. 5944–5958).

McMahon, A. M. (1994). *Understanding language change.* Cambridge university press.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., . . . others. (2011). Quantitative analysis of culture using millions of digitized books. *science, 331*, 176–182.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mitra, S., Mitra, R., Maity, S. K., Riedl, M., Biemann, C., Goyal, P., & Mukherjee, A. (2015). An automatic approach to identify word sense

changes in text media across timescales. *Natural Language Engineering, 21*, 773–798.

Mokhtar, K., Bukhari, S. S., & Dengel, A. (2018). OCR Error Correction: State-of-the-Art vs an NMT-based Approach. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, (s. 429–434).

Nguyen, T. T., Jatowt, A., Coustaty, M., & Doucet, A. (2021)**.** Survey of post-OCR processing approaches. *ACM Computing Surveys (CSUR), 54*, 1–37.

Onoe, Y., Zhang, M., Choi, E., & Durrett, G. (2022). Entity Cloze By Date: What LMs Know About Unseen Entities. *Findings of the Association for Computational Linguistics: NAACL 2022*, (s. 693–702).

Özavşar, R. (2013). Tarama Sözlüğü ve Türkçe Sözlüke göre anlam değişmeleri.

Öztürel, A., Kayadelen, T., & Demirsahin, I. (2019). A syntactically expressive morphological analyzer for Turkish. *Proceedings of the 14th international conference on finite-state methods and natural language processing*, (s. 65–75).

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one, 10*, e0137041.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora.

Rodríguez Guerra, A. (2016)**.** Dictionaries of Neologisms: a Review and Proposals for its Improvement. *Open Linguistics, 2*.

Rudolph, M., & Blei, D. (2018). Dynamic embeddings for language evolution. *Proceedings of the 2018 world wide web conference*, (s. 1003–1011).

Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D. R., & Tsvetkov, Y. (2020). Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. *arXiv preprint arXiv:2001.07740*.

Sagi, E., Kaufmann, S., & Clark, B. (2009). Semantic density analysis: Comparing word meaning across time and phonetic space. *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, (s. 104–111).

Sak, H., Güngör, T., & Saraçlar, M. (2008). Turkish language resources: Morphological parser, morphological disambiguator and web corpus. *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings*, (s. 417–427).

Salan, M., & KABADAYI, O. (2022). Çağdaş Türk yazı dillerinde art zamanlı söz başı ünlü düşmesi üzerine. *Türk Dili Araştırmaları Yıllığı-Belleten*, 85–107.

Say, B., Zeyrek, D., Oflazer, K., & Özge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. *Proceedings of the eleventh international conference of Turkish linguistics*, (s. 183–192).

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika, 31*, 1–10.

Sezer, B., Sezer, T., & Ünivesitesi, M. (2013). TS corpus: Herkes için Türkçe derlem. *Proceedings of the 27th national linguistics conference,* (s. 217–225).

Sharma, A., Chhablani, G., Pandey, H., & Patil, R. (2021, November). DRIFT: A Toolkit for Diachronic Analysis of Scientific Literature. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (s. 361–371). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. doi:10.18653/v1/2021.emnlp-demo.40

Soper, E., Fujimoto, S., & Yu, Y.-Y. (2021). BART for post-correction of OCR newspaper text. *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021),* (s. 284–290).

Stern, G. (1931). Meaning and change of meaning: With special reference to the English language. *Meaning and change of meaning: With special reference to the English language*. Indiana University Press.

Strötgen, J., & Gertz, M. (2012, May). Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. N. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, . . . S. Piperidis (Dü.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* içinde (s. 3746–3753). Istanbul: European Language Resource Association (ELRA).

Sultanzade, V. (2012). The syntactic valency of some verbs in The Book of Dede Korkut: diachronic differences. *bilig, 61*, 223.

Sulubacak, U., Eryiğit, G., & Pamay, T. (2016). IMST: A revisited Turkish dependency treebank. *Proceedings of TurCLing 2016, the 1st international conference on Turkic computational linguistics.*

Szymanski, T. (2017). Temporal word analogies: Identifying lexical replacement with diachronic word embeddings. *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: short papers),* (s. 448–453).

Tang, X., Qu, W., & Chen, X. (2016). Semantic change computation: A successive approach. *World Wide Web, 19*, 375–415.

Taylor, J. R. (2015). *The Oxford handbook of the word.* OUP Oxford.

TDK. (1935). *Türkçeden Osmanlıcaya Cep Kılavuzu.* Ankara: Turkish Language Association.

Thomas, A., Gaizauskas, R., & Lu, H. (2024). Leveraging LLMs for Post-OCR Correction of Historical Newspapers. *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024,* (s. 116–121).

Underhill, R. (1976). *Turkish grammar.* MIT press Cambridge, MA.

Ünal-Logacev, Ö., Żygis, M., & Fuchs, S. (2019). Phonetics and phonology of soft 'g' in Turkish. *Journal of the International Phonetic Association, 49*, 183–206. doi:10.1017/S0025100317000317

Vahit, T. Ü. (2003). Türkçede Ön Seste Ünlü Düşmesi Örnekleri. *Türk Dili Araştırmaları Yıllığı-Belleten, 49*, 223–233.

Xu, Y., & Kemp, C. (2015). A Computational Evaluation of Two Laws of Semantic Change. *CogSci.*

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. *Proceedings of the eleventh acm international conference on web search and data mining*, (s. 673–681).

Yavuzarslan, P. (2011). Türk dilinde kişi eklerinin tarihsel gelişimi ve değişimi. *38. ICANAS (10-15.09. 2007) Bildiriler (Dil Bilimi, Dil Bilgisi ve Dil Eğitimi)*, 1953–1966.

Yurddaş, R., & Karadağ, C. (2008). Tutanak Dergisi Münderecatı. *Yasama Dergisi*, 1–12.

Zhang, Y., Jatowt, A., & Tanaka, K. (2017). Temporal analog retrieval using transformation over dual hierarchical structures. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, (s. 717–726).

Zhang, Y., Jatowt, A., Bhowmick, S., & Tanaka, K. (2015). Omnia mutantur, nihil interit: Connecting past with present by finding corresponding terms across time. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, (s. 645–655).

Zheng, J., Ritter, A., & Xu, W. (2024). NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms. *arXiv preprint arXiv:2402.12261*.

URL-1 www.resmigazete.gov.tr, alındığı tarih: 16 Ekim 2024

URL-2 www.tbmm.gov.tr, alındığı tarih: 19 Ekim 2024

EKLER

EK 1: Lingan

EK 2: Bölüm 3’te kullanılan eş anlamlı kelimeler

EK 1

Bileşen	Metot	Açıklama
Data	*save(path)*	*Data* nesnesini, *path* tarafından belirtilen konuma kaydeder.
	load(path)	*Data* nesnesini belirtilen *path*'ten yükler.
	transform(f)	Ham veriyi *f* fonksiyonunu kullanarak dönüştürür.
	merge(other)	*other*'ın içeriğini *data* objesiyle ile birleştirir.
Container	*save(path)*	*Container* nesnesini belirtilen *path* konumuna kaydeder.
	load(path)	*Container* nesnesini belirtilen *path*'ten yükler.
	perform(Operation)	Bir *Operation* nesnesini parametre olarak alır ve *Operation* bileşeninde tanımlı ilgili yönlendirme metotlarını çağırır.
	period()	*Container* nesnesinin zaman aralığını döndürür.
	is_diachronic()	*Container* nesnesinin Diyakronik analiz için uygun olup olmadığını döndürür.
	is_synchronic()	*Container* nesnesinin Senkronik analiz için uygun olup olmadığını döndürür.

	add(cont)	*Container* nesnesi *cont*'u metodu çağıran *Container* yapısına ekler.
	remove(cont)	*Container* nesnesi *cont*'u metodu çağıran *Container* yapısından çıkarır.
	get(time_range)	Zaman dilimi *time_range* içine düşen tüm *Container* objelerini döndürür.
	get_data()	*Container* objesinde tutulan *Data* nesnesini döndürür.
Operation	*on_diachronic(DiachronicCorpus)*	Bir *DiachronicCorpus* nesnesini alır ve *DiachronicCorpus* üzerinde işlem mantığını uygular.
	on_synchronic(Corpus)	Bir *Corpus* nesnesini alır ve *Corpus* üzerinde işlem mantığını uygular.

İşlem	**Veriyapısı**	**Açıklama**
Exists(word, time_range)	*Vocabulary*	*word* kelimesinin *time_range* zaman aralığında bir mevcut olup olmadığını kontrol eder..
Frequency(word, time_range)	*Vocabulary*	*time_range* zaman aralığında, her bir bileşini *word* kelimesinin ilgili zaman dilimindeki frekansı olan bir zaman serisi döndürür.
FilterFrequency(threshold, time_range)	*Vocabulary*	*time_range* zaman aralığı içinde, frekansı belirtilen *threshold*

		değerinden daha düşük olan kelimeleri kaldırır.
VocabularySimilarity(time_range)	*Vocabulary*	Her bir elemanı farklı zaman dilimlerinde kullanılan kelimeler arasındaki *Jaccard Index* sonucunu temsil eden bir matris döndürür.
VocabularyDistance(time_range)	*Vocabulary*	Her bir elemanı farklı zaman dilimi ikilileri arasındaki JSD değerini temsil eden bir matris döndürür.
MorphemFrequency(morpheme, time_range)	*Vocabulary*	*time_range* zaman aralığında *morpheme*'in kullanım frekansını döndürür.
UniqueWordCount(time_range)	*Vocabulary*	Her bir elemanı, *time_range* aralığında bulunan benzersiz kelimelerin toplam sayısını temsil eden bir zaman serisi döndürür.
CommonWords(time_range)	*Vocabulary*	*time_range* içerisinde bulunan ve her korpusta var olan kelimelerden oluşan bir küme döndürür
CoFrequency(u, v, time_range)	*Embeddings*	Zaman aralığı *time_range* içinde, *u* ve *v* kelimelerinin birlikte görülme frekansını temsil eden bir zaman serisi döndürür.
Collocations(word, k, time_range)	*Embeddings*	Her bir elemanı bir kelime kümesi olan bir dizi döndürür. Büyüklüğü *k* olan bu kümeler, her zaman

		dilimi için *word* ile en yüksek ortak kullanım değerine sahip kelimelere karşılık gelir.
Association(u, v, time_range)	*Embeddings*	Her bir elemanı ilgili zaman diliminde *u* ve *v* kelimeleri arasındaki ortak kullanım değerine karşılık gelen bir zaman serisi dizisi döndürür.
Similarity(u, v, time_range)	*Embeddings*	Her bir elemanı ilgili zaman diliminde *u* ve *v* arasındaki semantik benzerliğe karşılık gelen bir zaman serisi döndürür.
AlignedMostSimilar(word, k, target_period, base_period)	*Embeddings*	*target_period*'un vektör uzayını *base_period'*ın vektör uzayına hizalar ve *base_period* zaman diliminden w*ord* kelimesine semantik olarak en benzer *k* kelimeyi döndürür
SemanticChange(word, time_range)	*Embeddings*	Her bir elemanı, w*ord* kelimesinin başlangıç dönemindeki vektöründen olan uzaklığı temsil eden bir zaman serisi dizisi döndürür.

EK 2

bakan - vekil
başkan - reis
görev - vazife
tl - lira
milletvekili - mebus
kurul - heyet

uygula - tatbik
önerge - takrir
üye - aza
yıl - sene
tasarı - lâyiha
genel - umum
yasa - kanun
ülke - memleket
işlem - muamele
sun - takdim
gelir - varidat
oy - rey
okul - mektep
süre - müddet
özel - hususî
durum - vaziyet
yardım - muavenet
toplam - yekûn
yabancı - ecnebi
sağlık - sıhhat
ilgi - alâka
tüzük - nizamname
başbakan - başvekil
yetki - salâhiyet
ek - ilâve
tekel - inhisar
uygun - muvafık
belge - vesika
önce - evvel
etki - tesir
tarım - ziraat
incele - tetkik
son - nihayet
neden - sebeb
sonuç - netice
yön - cihet
ilçe - kaza
seçim - intihap
isim - ad
yolluk - harcırah
karşılık - mukabil
ilişkin - müteallik
yer - mahal
gündem - ruzname
biçim - tarz
adet - aded
satınal - mubayaa
gerek - icap
olağanüstü - fevkalâde
istek - talep

uzman - mütehassıs
toplantı - içtima
belirli - muayyen
sekreter - kâtip
yürürlük - meriyet
göre - nazaran
kullan - istimal
denet - murakabe
konu - mevzu
uluslararası - beynelmilel
dernek - cemiyet
şart - şerait
kişi - şahıs
yarar - istifade
değer - kıymet
yayım - neşir
sorun - mesele
tutanak - zabıt
önem - ehemmiyet
araç - vasıta
aralık - kânunuevvel
yeter - kâfi
nitelik - mahiyet
onarım - tamirat
dağıtım - tevzi
indirim - tenzilât
harca - sarfiyat
öğrenci - talebe
tüketim - istihlâk
gazete - ceride
dilekçe - istida
iç - dahil
savaş - harp
onay - tasdik
kuzey - şimali
ocak - kânunusani
güney - cenup
sayman - muhasip
yüküm - mükellef
geliş - inkişaf
saklı - mahfuz
gerçek - hakikî
kapı - bap
yaşlı - ihtiyar
yayın - neşriyat
parti - fırka
ekim - teşrinievvel
değişik - muaddel
bölüm - kısım
özellikle - bilhassa

kasım - teşrinisani
basın - matbuat
üretim - istihsal
cevap - cevab
engel - mâni
kağıt - kâğıd
yasak - memnu
cumhurbaşkanı - reisicumhur
kayıt - kayıd
başarı - muvaffakiyet
sorumlu - mesul
tutar - meblağ
konut - mesken
üniversite - darülfünun
koru - muhafaza
aday - namzet
suç - cürüm
aykırı - hilâf
yüce - ulvi
bildirim - beyanname
il - vilâyet
disiplin - inzibat
doğu - şark
benzer - mümasil
öğretim - tedrisat
sınav - imtihan
olay - hâdise
oysa - halbuki
yakacak - mahrukat
ortak - şerik
emekli - mütekait
soruştur - tahkikat
dış - haricî
ekonomi - iktisat
üst - mafevk
kapsam - şümul
çoğunluk - ekseriyet
örnek - nüsha
zorunlu - mecburi
vade - mühlet
çıkar - menfaat
eksik - noksan
yasak - memnuiyet
ama - amma
tutuk - mahkûm
kapsa - ihtiva
başlangıç - bidayet
kural - kaide
sakat - malûl
kötü - fena

ödül - mükâfat
çoğunlukla - ekseriyetle
kesin - katî
ilgili - alakadar
alkol - ispirto
fiyat - fiat
saldırı - taarruz
belirti - alâmet
tören - merasim
gerçek - hakikî
oluş - vuku
eşdeğer - muadil
dil - lisan
ekonomik - iktisadî
çocuk - evlâd
matematik - riyaziye
kaynak - memba
temsil - mümessil
gözlem - rasat
çok - pek
sahip - malik
gözetim - nezaret
hüküm - ahkâm
ortalama - vasati
ilişki - münasebet
profesör - müderris
fiilen - bilfiil
koca - zevç
çizgi - hat
artma - tezayüt
arama - taharri
bayındır - nafıa
atölye - atelye
boş - münhal
yük - hamule
gezi - seyahat
salt - mutlak
amaç - maksat
kavram - mefhum
birçok - müteaddit
batı - garp
müzik - musiki
eşit - müsavat
zor - müşkül
kira - isticar
bitki - nebat
az - kıt
rast - tesadüf
dayan - istinaden
çevre - muhit

varlık - mevcudiyet
güven - itimat
kesim - mukataa
kimlik - hüviyet
yoksun - mahrum
kamu - âmme
tüzel - hükmî
doğum - tevellüt
albay - miralay
yasama - teşrî
tanık - şahid
belediye - şehremaneti
gösterge - icmal
zarar - ziya