# 25. Lematizasyon ve Türkçe için bir lematizasyon uygulaması: elemanTR

**Bekir Tahir TAHİROĞLU** [1]

**APA:** Tahiroğlu, B. T. (2021). Lematizasyon ve Türkçe için bir lematizasyon uygulaması: elemanTR. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, (24), 475-486. DOI: 10.29000/rumelide.995343.

## Öz

Madde başı (lemma) bir dildeki sözcüklerin sözlüklerde yer alan standart biçimi olduğu bilinmektedir. Lematizasyon adı verilen süreç ise çekimli sözcüklerin madde başına indirgeme sürecidir. Lematizasyon çeşitli diller için doğal dil işleme (DDİ) araçlarında metinlerin ön işleme aşamasında sözcüklerin farklı yapılarının normal biçimlerine dönüştürülmesinde kullanılmakla birlikte, dilbilimsel açıdan sözcük ailelerinin otomatik oluşturulması ve özellikle sözlüklerin hazırlanmasında büyük kolaylıklar sağlamaktadır. Milyon sözcüklük bir derlemdeki tüm sözcüklerin madde başı biçimlerine otomatik ve doğru biçimde indirgenmesi hem zaman hem de emek yoğun işlemlerin verimli olarak yapılmasını sağlamaktadır. Lematizasyon işlemi, sözcüklerin kullanım sıklığına bağlı çözümlemelerde de çekimli biçimlerin ayrı sayımından kaynaklanan zaman kaybını da gidererek ilgilenilen metnin veya derlemin anahtar kavramlarını çok kısa sürede ortaya çıkarmaktadır. Bu çalışmada, genel olarak lematizasyon üzerinde durulmuş ve Türkçe için bağlam duyarlı olarak hazırlanan ve elemanTR adı verilen bir lematizasyon uygulama modeli tanıtılmıştır. Türkçe 184 adet roman ve hikaye metinlerinden hazırlanan yaklaşık 1 milyon 200 bin cümleyle makine öğrenmesiyle eğitilen modelde elde edilen başarım test verilerinde %99, 01 olarak bulunmuştur. Elde edilen bu başarım oranının eğitim verisine yeni eklenecek verilerle sürdürülebilir bir yapıya kavuşturularak genel bir Türkçe lematizasyon uygulamasının ileri çalışmalar için ön ayak olacağı düşünülmektedir. Otomatik söz türü belirleme, sözdizimsel çözümleme ve yeni sözcüklerin otomatik belirlenmesinde de eğitilen modelin yararlı olacağı ön görülmektedir.

**Anahtar kelimeler**: Lema, lematizasyon, söz varlığı, doğal dil işleme, yapay öğrenme

# Lemmatization and a lemmatization application for Turkish: elemanTR

## Abstract

Lemma is the standard form of words found in dictionaries in a language. The process called lemmatization is the reduction process of inflected words. Although lemmatization is used in natural language processing (NLP) tools for various languages to convert different structures of words into standard forms during the preprocessing stage of texts, it provides great convenience in the automatic creation of word families, especially in the preparation of dictionaries in terms of linguistics. Automatic and accurate reduction of all words in a million-word corpus to lemma forms provides efficient execution of both time- and labor-intensive processes. The lemmatization process reveals the key concepts of the text or corpus of interest in a very short time by eliminating the loss of time caused by the separate counting of inflected forms in the frequency analysis of words. In this study, lemmatization has been focused on in general, and a context-sensitive lemmatization application model called elementTR has been introduced for Turkish. The performance of the model, which was

[1] Dr. Öğr. Üyesi, Çukurova Üniversitesi, Fen Edebiyat Fakültesi, Türk Dili ve Edebiyatı Bölümü (Adana, Türkiye), tahirbekir@gmail.com, ORCID ID: 0000-0002-7956-3257 [Araştırma makalesi, Makale kayıt tarihi: 16.08.2021-kabul tarihi: 20.09.2021; DOI: 10.29000/rumelide.995343]

trained with machine learning with approximately 1 million 200 thousand sentences prepared from 184 novels and story texts in Turkish, was found to be 99.01% in the test data. It is thought that this success rate will be a sustainable structure with the new data to be added to the training data, and a general Turkish lemmatization application of this model will be a pioneer for further studies. It is predicted that the trained model will also be useful in automatic part of speech identification, syntactic analysis, and automatic identification of new words.

**Keywords**: Lemma, lemmatization, vocabulary, natural language processing, machine learning

## 1. Giriş

Derin öğrenme ve doğal dil işleme (DDİ) araştırmalarında dil verisinin kullanımı giderek artmakta, veriye dayalı uygulamalar da buna koşut olarak çeşitlenmektedir. DDİ'de son yıllarda beynin çalışma biçiminden hareketle oluşturulan yapay sinir ağlarına dayalı uygulamaların sayısı ve buna bağlı olarak çeşitli düzeyde geliştirilen dil çözümleme araçlarının başarımı da artmıştır. Yapay zekanın bir çalışma alanı olan makine öğrenmesine dayanan uygulamalar, büyük verinin insan dilinin gerek biçimsel ve sözdizimsel gerekse anlamsal yapısının birçok yönden ele alınmasını sağlayan yeni yöntemlerin geliştirilmesine de öncülük etmektedir.

Yapay öğrenme olarak da adlandırılan makine öğrenmesini Alpaydın, bilgisayarların veriden öğrenerek başarımlarını artıracak biçimde programlanması olarak tanımlamaktadır. Yapay öğrenme bir bakıma, veriden istatistik ve olasılık hesapları kullanılarak bir model çıkarma işidir. Veriden öğrenen makine, veriye dayalı özellikleri bir model biçiminde defalarca kullanılacak ve genel ya da özel bir problemi çözecek formata dönüştürmektedir. Yapay öğrenmenin amacı makinenin probleme dönük eğitimi sırasında belirlenen parametre değerlerini iyileştirmektir (Alpaydın, 2011, s.3).

Dil işlemede makine öğrenmesi yöntemlerinin yaygınlaşması tokenizasyon, söz türü etiketleme, lematizasyon gibi bir dil derleminin işlenmesinde veri ön işleme aşamalarında (normalizasyon) yeni arayışları gündeme getirmiştir. Kural tabanlı ve istatistiksel biçibirimsel ve sözdizimi çözümleme çalışmalarında kural tabanlı yöntemlerden sonra istatistiksel yöntemler denenmiş fakat bunların başarım oranları sınırlı kalmıştır. Bu nedenle yapay öğrenme teknikleri günümüzde standart duruma gelmiştir denilebilir. Lematizasyon gibi veri ön işleme araştırmalarında da yeni yöntemler üzerinde durulmakta, açık kaynaklı olarak birçok proje çeşitli internet platformlarında yerini almaktadır.

Lematizasyon (*lemmatization*) temel olarak sözcüklerin çekimli biçimlerini "otomatik" olarak sözlükte geçen madde başı biçimine dönüştürme/indirgeme işlemi ve bir tür sınıflandırmadır. Lema (*lemma*) sözcüğün sözlükte yer alan madde başı biçimidir. İnsanın manuel olarak bir sayfa metinde lematizasyonu yapması mümkünken binlerce sayfa metinde belirli bir zaman aralığında (saniyeler içinde) bu işlemi doğru biçimde yapabilmesi mümkün değildir. Metinlerin elektronik olduğu ve üretildiği bu çağda gerek üslup çalışmaları gerek sözlük hazırlama gerekse DDİ'de yapay zekâ uygulamaları için bir ön işlem durumunda yer alan lematizasyon önemli bir aşamaya gelmiştir. Lematizasyonla birlikte söz varlığı araştırmalarında çekimli biçimlerle birlikte sözcük ailelerinin otomatik tespit edilmesi de mümkün olabilmektedir.

Bu çalışmanın konusu olan lematizasyon, basitçe dildeki çekimli biçimlerin dilin standart olarak kabul edilen sözlüğünde yer alan madde başı formuna otomatik olarak dönüştürülmesi işlemidir. Lematizasyon sözcüğünün Türkçedeki terim karşılığı üzerinde henüz bir birlik sağlandığını

söyleyemeyiz. Bununla birlikte kavram karşılığı olarak *sözlükbirimleştirme* olarak da adlandırılabilmektedir.2 Lematizasyonda amaç uygulamaya girdi olarak sunulan işlenmemiş bir metni ya da derlemi insan müdahalesiz çekimli biçimlerden sözlükbirimler biçimine dönüştürmektir. Burada yapılan işlem bağlama bağlı ya da bağlamdan bağımsız olarak yapılabilir. Bağlamdan bağımsız yapılan lematizasyonda uygulama eş sesli sözcüklerin ortaya çıkardığı belirsizlik durumunu çözememektedir. Örneğin "yer" sözcüğünün isim ya da fiil biçiminde kullanım farkından kaynaklı belirsizlik problemini çözemez, "bu ekmeği yerim" cümlesinde "yerim" sözcüğünü [yemek] biçimine dönüştürmesi gerekirken [yer] madde başına dönüştürebilir.

Lematizasyon uygulamaları birçok dil için geliştirilmiş, Türkçe DDİ çalışmalarında da günümüzde çalışılan bir konu duruma gelmiştir. Türkçenin sondan eklemeli yapısı dolayısıyla zengin bir biçimbirimsel yapısının olması, Hint-Avrupa dil ailesinden olan dil işleme açısından çok çalışılan bir dil olan İngilizceye göre lematizasyon probleminin çözümünde zor bir durumda olmasına neden olmaktadır. Özellikle çekim eklerindeki çeşitliliğin çokluğu, morfotaktik olarak ek sıralanmasının belirli bir kurala bağlı olmasına karşın eklerin sözcük gövdelerine eklenmesi sırasındaki değişimler de lematizasyon işlemini zorlaştıran unsurlar arasındadır.

LEMMING3 adı verilen ve altı dil (İngilizce, Almanca, Çekçe, Macarca, Latince ve İspanyolca) için geliştirilen lematizasyon sisteminde joint log-linear ve token tabanlı bir yöntem kullanılarak bilinmeyen sözcüklerin lematizasyonu problemi çözülmeye çalışılmıştır. Sistemde biçimbirimsel sözlüklük yardımı kullanılmamıştır. Yani lematizasyonu yapılacak birimler için bizim çalışmamızda olduğu gibi doğru lematizasyon biçimini tahminleyecek yardımcı bir özellik bulunmamaktadır. Çalışmada lemaların kendisi bir özellik olarak ele alınmış, lemalardaki ön ve son eklerin sınırları 10 karakter uzunluğu olarak belirlenmiştir. Çıktı olarak sözcük türlerinin de işaretlendiği çalışmada tüm diller için 100 bin token kullanılmıştır. Sözü edilen diller için global bir sistem geliştirilmeye çalışılmış ve sonuçta sözcük türü bilgisinin başarımı artırdığı bulunmuştur. LEMMING, Çekçe için %96,29; Almanca için %97,84; İngilizce için %98,71; İspanyolca için %97,91; Macarca için %97,31 ve Latince için %93 oranında başarım göstermiştir. Bu oranlar sözlük ve biçimbirimsel özellik yardımlarıyla %98'lik başarım oranlarına ulaşmıştır (Müller vd., 2015, s. 2268-2272).

Türkçeye özgü lematizasyon uygulamaları geliştirilmeye çalışılmıştır. Öztürkmenoğlu ve Alpkoçak, Türkçe bilgi erişimi (*information retrieval*) için bir metin koleksiyonu üzerinde farklı lematizasyon yaklaşımlarını incelemişlerdir. Sözlük tabanlı (DTL), son durumlu makine (Oflazer's Morphological Analyzer (OMA) ve sabit karakter uzunluğu yaklaşımlarını ve açık kaynak kodlu Zemberek uygulamasının sonuçlarını Bilken Milliyet Koleksiyonu üzerinde karşılaştırmışlar ve sözlük tabanlı (TDK Büyük Tükçe Sözlük'ün kullanılmıştır) ağaç yapılı algoritmanın daha başarılı sonuçlar üreterek Türkçe bilgi erişim sisteminin daha kullanışlı olduğunu bulmuşlardır (Öztürkmenoğlu ve Alpkoçak, 2012, s.2-3) .

Türkçe lematizasyon uygulaması için Arslan ve Orhan biçimbilgisel benzerlikten yararlanılarak çizge tabanlı bir model önermişlerdir. LemmaTuR adını verdikleri modelde bir bilgisayar bilimlerinde çizge ya da graf veri yapısı olarak bilinen bir yapıya tüm sözcükleri düğüm (node) olarak kabul etmişler ve bu düğümlerin ilişkilerini benzerlik fonksiyonlarıyla hesaplayarak bir graf ağı oluşturmuşlardır. Modelde, Türkçenin sondan eklemeli türetme yapısında sözcük gövdesine gelen çekimli biçimlerin gramatikal olarak benzer komşuluk ilişkisine sahip sözcükleri kapsadığı varsayımından hareket edilmiştir. Bu

2 http://turkcederlem.mersin.edu.tr/okulsozluk/sayfa.php?s=amac&dil=tr
3 http://cistern.cis.lmu.de/lemming/

şekilde hazırlanan modelin %95,9 oranında doğru madde başı tahmini yaptığı belirlenmiştir (Arslan ve Orhan, 2016, s.2-4).

Bergmanis ve Goldwater tarafından farklı diller için yapay sinir ağı modeliyle bağlam duyarlı biçimde hazırlanan ve Lematus adı verilen sistem aralarında Türkçenin de bulunduğu 20 dil üzerinde denenmiştir. Her dil için 10 bin eğitim örneği ile hazırlanan sistemde Türkçe için %85,9'luk bir başarı oranı elde edilmiştir (Bergmanis ve Goldwater, 2018, s. 1391,1396).

Lematizasyon sadece DDİ araçlarının geliştirilmesinde değil dilbilimsel incelemelerde de kolaylıklar sağlamaktadır. Sözlüklerin hazırlanması sırasında madde başlarının otomatik tespiti, sözcük sıklıklarının çözümlemesinde çekimli biçimlerin kısa sürede madde başlarına indirgenerek anahtar kavramların bulunmasında da kolaylıklar sağlamaktadır. Üslup araştırmaları ya da yazar sözlüklerinin oluşturulması, yazara ait gramatikal dizin ve sözcük ailelerinin belirlenmesi de otomatik gerçekleştirilen lematizasyonun sağladığı yararlar arasındadır.

## 2. Yöntem

### 2.1. Veri

Otomatik lematizasyon yazılımının oluşturulmasında kullanılan veri Türkçe roman ve hikâye metinlerine dayanmaktadır. Bu bakımdan elde edilen başarımda, girilen sözcüğün ve bağlam özelliklerinin roman ve hikâye özelliklerine bağlı olabileceği göz önünde bulundurulmuştur.

Çalışmada, eğitim ve test setleri için kullanılan derlem roman ve öykü metinlerinden yararlanılmıştır. Derlemin büyüklüğü için hedef bir sözcük ve metin sayısı önceden planlanmamış, verideki sözcük sayısı büyüdükçe elde edilecek başarımdaki (performans) doğruluğun yüksek olacağı öngörüsüyle hareket edilmiştir. Eğitim (train), gelişme (devolopment) ve test verisi olarak üç bölüme ayrılan veri setlerindeki sözcükler satırlaştırılarak iki sütunlu biçimde saklanmıştır. Birinci sütunda sözcüğün kaynakta geçtiği biçimi (çekimli ya da madde başı biçimi), ikinci sütundaysa her bir sözcüğün lema ya da maddebaşı biçimi yer almaktadır. Aşağıdaki şekilde eğitim verisinde kullanılan yapıdan bir örnek yer almaktadır.

```
12718442  Güneş⟶lema
12718443  henüz⟶lema
12718444  o⟶lema
12718445  hizaya⟶hiza
12718446  kadar⟶lema
12718447  yükselmemişti⟶yükselmek
12718448  ve⟶lema
12718449  yarları⟶yar
12718450  dolduran⟶doldurmak
12718451  dumanlar⟶duman
12718452  iyi⟶lema
12718453  seçmesini⟶seçme
12718454  engelliyordu⟶engellemek
12718455  .⟶.
12718456
12718457  Yine⟶lema
12718458  de⟶lema
12718459  Drogo⟶oa
12718460  hızlanarak⟶hızlanmak
12718461  o⟶lema
12718462  şeyle⟶şey
12718463  aynı⟶lema
12718464  hizaya⟶hiza
12718465  gelmeyi⟶gelme
12718466  başardı⟶başarmak
12718467  ve⟶lema
12718468  bunun⟶bu
12718469  bir⟶lema
12718470  adam⟶lema
12718471  ,⟶,
```

Şekil 1. elemanTR eğitim ve gelişme verisi yapısı.

Şekil 1'de satır numaralarından sonra yer alan iki sütun esas yapıyı oluşturmakta ve sol sütundaki sözcükler bir tab karakteri ile sözcüğün madde başından (lema) ayrılarak kullanılan yapay öğrenme sistemine girdi olmaktadır. Şekildeki yapı, gelişme veri setinde de kullanılmıştır. Test veri setinde sadece soldaki giriş sütunu sistemin tahminleme bileşenine girdi olmaktadır. Her sözcük satırlaştırılmış (tokenizasyon) ve sözcüklerdeki kesme işaretleri dâhil tüm noktalama işaretleri korunmuştur. Noktalamaya dayalı cümle sınırları (., ?, !, :, …) belirlenerek bu sınırlardan sonra bir satır boşluk ayracı kullanılarak cümlelere ayrılan yapı oluşturularak sistemin kabul ettiği biçime dönüştürülmüştür. Tüm bunlar çalışmanın veri ön işleme aşamalarını oluşturmaktadır. Sözcüklerin türleri etiketlenmemiş sadece çıktı durumları (lema) hedef durum olarak sisteme gösterilmiştir. Özel adlar "OA" etiketiyle temsil edilmiş, zaten madde başı olan sözcükler "lema" olarak etiketlenmiştir. Bu şekilde lema biçiminde etiketlenen sözcüklerin madde başı formlarının olduğu gibi kullanılmamasındaki amaç sisteme girdi olacak sınıf sayısının düşürülerek başarımın artırılmasıdır.

Veri setinin hazırlanmasında kullanılan toplam roman ve hikâye sayısı 193'tür. Eğitim setinde 1.169.694, gelişme setinde 19.969 ve test setinde 25.025 olmak üzere tüm veri setinde 1.214. 688 cümle kullanılmıştır. Cümlelere ayrılmamış veri, toplam 13.740.541 satırdan oluşmaktadır yani bu,

uygulamanın eğitilmesinde kullanılan toplam birim (*token*) sayısıdır. Verideki farklı sözcük sayısı (*type*) ise 462.882'dir. Sözcük çeşitliğini gösteren type/token oranı ise %67,10'dur. Çalışmada sözcüklerin kök ve etimolojik özellikleriyle sözcük türleri destekleyici olarak kullanılmamıştır. Sözcüklerin çekimli biçimleriyle lemaların kendileri kullanılan uygulamanın girdi olarak aldığı özellikler olmuştur.

### 2. 2. Kullanılan teknoloji

Çalışmada yapay sinir ağlarının son yıllarda geliştirilen ve derin öğrenme (*deep learning*) adı verilen teknolojisine dayalı açık kaynak kodlu olarak hazırlanan Ubiquitous Knowledge Processing Lab tarafından geliştirilen "emnlp2017-bilstm-cnn-crf"4 adlı platform kullanılmıştır. Kullanılan bu platform DDİ çalışmalarında söz türü etiketleme (*part of speech tagging*) olarak adlandırılan ve bir metinde ya da derlemdeki tüm sözcüklerin türlerinin insan müdahalesiz otomatik biçimde belirlenip etiketlenmesini sağlayan bir amaç için geliştirilmiştir. ElemanTR uygulamamızda lematizasyon amaçlı olarak bu etiketleme sistemi (söz türü etiketleme) girdi yapısına uygun olarak platformun varsayılan parametre ayarlarının yeninden düzenlenmesi şeklinde kullanılmıştır. Eğitim sırasında yaklaşık 20 bin satırlık ilk veri seti manuel olarak girdi->lematizasyonlu form ikilisi şeklinde işlendikten sonra makinenin eğitimine geçilmiş ve bir sonraki her 20 bin satırlık veri, öğrenme sürecinde sonuç olarak alınarak eğitim seti sürekli geliştirilmiştir. Bu yolla, her eğitim zamanında makine daha önce karşılaşmadığı sözcüklerle eğitilerek hedeflenen en yüksek başarım oranına ulaşılmıştır.

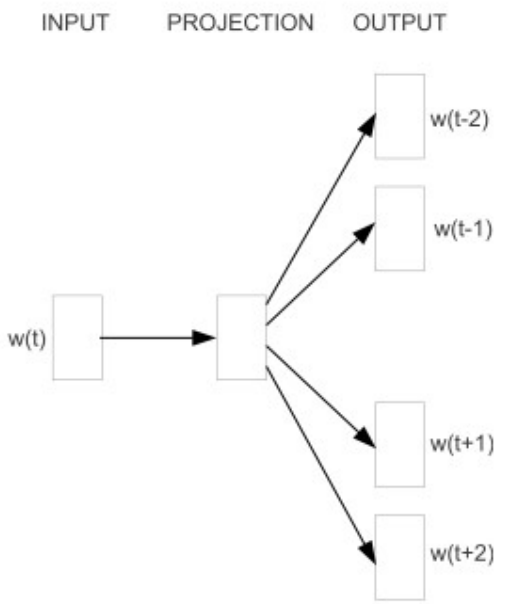
Yapay sinir ağlarına (nöral networks) dayanan makine öğrenmesi modellerinden olan LSTM modeli derin öğrenme yaklaşımlarından biridir. Geri dönüşlü yapay sinir ağları (RNN-Recurrent Neural Network) son yıllarda geliştirilmiş ve dil gibi sıra dizgesel yapıları (sözcük ve harflerin dizi biçimde olması) modellemek için kullanılmıştır. Bu ağ yapısındaki kimi sorunları aşmak için ise yine RNN mimarisi tabanlı LSTM ağları geliştirilerek sıra dizgesel yapının modellenmesi güçlendirilmiştir (Özkan, 2021, s.159, 172).

emnlp2017-bilstm-cnn-crf platformu, python diliyle yazılan bir platformdur. Derin öğrenmede dil işleme araştırmalarında başarımı yüksek kabul edilen ve bir dizinin (*squence*) birimlerinin öncelik sonralık sırasını uzun dönemli belleğinde tutabilen LSTM (*Long short-term memory*) tekniğini desteklemektedir. İki taraflı (*Bİ-LSTM*) olarak desteklenen bu teknikte girdi olarak alınan cümledeki her sözcüğün sağında ve solunda kalan sözcüklerin özellikleri hesaplanarak bağlam duyarlılığı kazandırılmaya çalışılmaktadır.

Sözcüklerin vektörlere dönüştürülmesi, sözcüklerin bağlamının tahmin edilmesinde günümüzde standart duruma gelmektedir. Mikolov vd. birçok DDİ sisteminde istatistiksel yöntemlerin kullanıldığını, bunlardan n-gram (sözcük dizilerindeki olasılıksal birliktelikler) temelli tekniklerin örneğin makineli dil çeviri sistemlerinde milyar sözcüklük bir veriyi gerektirdiğinden söz ederek daha yenilikçi ve ölçeklenebilir yöntemlere ihtiyaç duyulduğunu belirtmiştir. Bu durumda sözcük bağlamı tahminlemesinde matematiksel bir temsil olan ve lineer cebir alanında da yoğun biçimde kullanılan vektör uzaylarının yararlı olacağını öngörmüşlerdir. Çalışmalarında sözcüklerin dağılımsal temsili ile elde edilen skip-gram modeli önerilmiş ve formülleştirilmiştir. Skip-gram modelinde aşağıdaki şekilde de görüleceği gibi her sözcük iki birim öncesi ve sonrası düzeyinde tahminlenmektedir. Yani sözcüğün kendisi, sözcüğün bağlamını vermektedir. Sözcükten bağlama giden bir tahminleme söz konusudur

4 https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

(Mikolov vd., 2013, s.1) . elemanTR uygulaması için platformun varsayılan vektör boyutu olan 100 boyut ayarı kullanılmıştır.




Şekil 2. Sözcüklerin skip-gram modeli yapısı (Mikolov vd., 2013, s.5)

Şekil 2'de girdi (input) ve projeksiyon (projection) ve çıktı (output) olarak üç yapı yer almakta, her sözcüğe bağlı sözcük bağlamı [w (t-2) iki öncesi, w(t+2) iki sonrası] biçiminde dağılımsal yapıyı oluşturmaktadır. Kısaca bu modelde sözcük verildiğinde onun bağlamı tahmin edilmektedir. Bu yapının genel olarak büyük boyutlu veride daha başarılı olduğu belirtilmektedir.

Bu çalışmada, sözcük vektörü oluşturulurken de skip-gram tekniği kullanılmıştır. Çalışma verisindeki tüm sözcük ve noktalama özellikleri de yine açık kaynak kodlu word2vec[5] uygulamasıyla sözcük vektörlerine dönüştürülmüş, dönüştürülen bu yapı emnlp2017-bilstm-cnn-crf platformunda başarımı yükseltmek ve sözcük bağlamı özelliklerini artırmak üzere ayrı bir katman olarak kullanılmıştır. Olasılık kuramına bağlı tekniklerle geliştirilen (Markov modeli gibi) sözcük türü etiketleme araçları yerine başarımı daha yüksek ve sözcük bağlamının daha güçlü temsil edildiği vektör tabanlı yöntem seçilerek özellikle bağlam belirsizliği giderilmeye çalışılmıştır. Böylece hangi bağlamda hangi sözcük hangi bağlam durumunda hedeflenen çıktı durumunu alabilirin yolu daha olanaklı duruma gelmiştir. Bu durum özellikle "kazan" veya "koyun" gibi hedef çıktı durumu kazanmak, kazmak fiilleri ve kazan ismi; koy ismi, koymak fiili, koyun hayvan ya da kucak ismi olma durumlarıyla baş edebilmekte yararlı olmaktadır.

5 https://github.com/danielfrg/word2vec

## 3. Bulgular

elamanTR uygulamasının eğitiminden elde edilen sonuçlara göre 30 iterastonla yapılan denemede %99,01 test verisi doğruluğu elde edilmiştir. Bu doğruluk oranı bizim verimizde ulaştığımız en yüksek doğruluk oranı olup uygulamanın eğitim verisinin büyüklüğüne bağlı olarak yükselmesi olasıdır.

| | |
|---|---|
| 1 | 1unidep_pos0.98604212566432450.98871899724419950.98604212566432450.9887189972441995 |
| 2 | 2unidep_pos0.98651578951942680.98980353809227480.98651578951942680.9898035380922748 |
| 3 | 3unidep_pos0.98697207121470880.98964352386878830.98697207121470880.9896435238687883 |
| 4 | 4unidep_pos0.98738924305039520.98973686549915550.98738924305039520.9897368654991555 |
| 5 | 5 unidep_pos 0.9874109707501706 0.9901191216997066 0.9874109707501706 0.99011912 |
| 6 | 6unidep_pos0.98710678295331590.98973242065961420.98741097075017060.9901191216997066 |
| 7 | 7unidep_pos0.98728495009147370.98992799359943110.98741097075017060.9901191216997066 |
| 8 | 8unidep_pos0.98721107591223740.98991465908080720.98741097075017060.9901191216997066 |
| 9 | 9unidep_pos0.9871806571325520.98982576228998130.98741097075017060.9901191216997066 |
| 10 | 10unidep_pos0.98735013319079970.98983909680860520.98741097075017060.9901191216997066 |
| 11 | 11unidep_pos0.98699379891448410.98983465196906400.98741097075017060.9901191216997066 |
| 12 | 12unidep_pos0.98718934821246220.98978575873410970.98741097075017060.9901191216997066 |
| 13 | 13unidep_pos0.98690254257542780.98976353453640320.98741097075017060.9901191216997066 |
| 14 | 14unidep_pos0.98709809187340570.98971019646190770.98741097075017060.9901191216997066 |
| 15 | 15unidep_pos0.98696338013479870.98960796515245800.98741097075017060.9901191216997066 |
| 16 | 16unidep_pos0.98661573693839330.98988354520401810.98741097075017060.9901191216997066 |
| 17 | 17unidep_pos0.98680694069641630.98954129255933860.98741097075017060.9901191216997066 |
| 18 | 18unidep_pos0.98693296135511320.98957685127566900.98741097075017060.9901191216997066 |
| 19 | 19unidep_pos0.98683301393614660.98965241354787090.98741097075017060.9901191216997066 |
| 20 | 20unidep_pos0.98656359045893250.98932349542181530.98741097075017060.9901191216997066 |
| 21 | 21unidep_pos0.98663311909821350.98947906480576050.98741097075017060.9901191216997066 |
| 22 | 22unidep_pos0.98667222895780910.98948350964530190.98741097075017060.9901191216997066 |
| 23 | 23unidep_pos0.98645929750001080.98923015379144810.98741097075017060.9901191216997066 |
| 24 | 24unidep_pos0.98648102519978620.98965241354787090.98741097075017060.9901191216997066 |
| 25 | 25unidep_pos0.98649406181965140.98932794026135660.98741097075017060.9901191216997066 |
| 26 | 26unidep_pos0.98635065900113420.98931016090319130.98741097075017060.9901191216997066 |
| 27 | 27unidep_pos0.98649406181965140.98923015379144810.98741097075017060.9901191216997066 |
| 28 | 28unidep_pos0.98615510970315610.98923904347053070.98741097075017060.9901191216997066 |
| 29 | 29unidep_pos0.98650709843951660.98936349897768690.98741097075017060.9901191216997066 |
| 30 | 30unidep_pos0.98650709843951660.98936794381722810.98741097075017060.9901191216997066 |

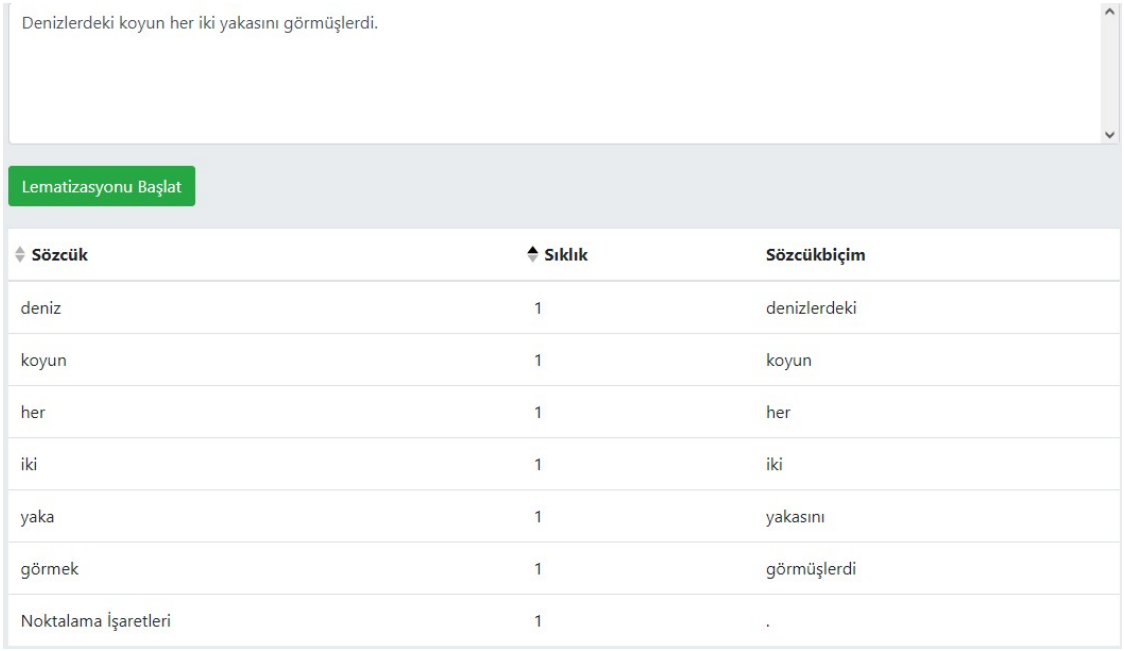
Şekil 3. elamanTR 30 iterasyon ile elde edilen skorlar.

Şekil 3'te 5. iterasyondan itibaren hem gelişme hem de test sonuçlarının en yüksek sonuçlara ulaştırğı görülmektedir. Daha yüksek skor elde edilmesi amacıyla 30 iterasyona kadar eğitim sürdürülmüş ve daha yüksek skorlara ulaşılmayınca eğitim sonlandırılmıştır.

Aşağıda örnek cümleler içinde kullanılan ve belirsizlik içeren cümlelere ait sonuçlar yer almaktadır.

Örnek Cümle 1:

"Denizlerdeki **koyun** her iki yakasını görmüşlerdi." Cümlede koyun sözcüğü belirsizlik içeren bir sözcüktür. Hedef çıktı koy+un "denizde bulunan bölüm" anlamında koy ismidir.

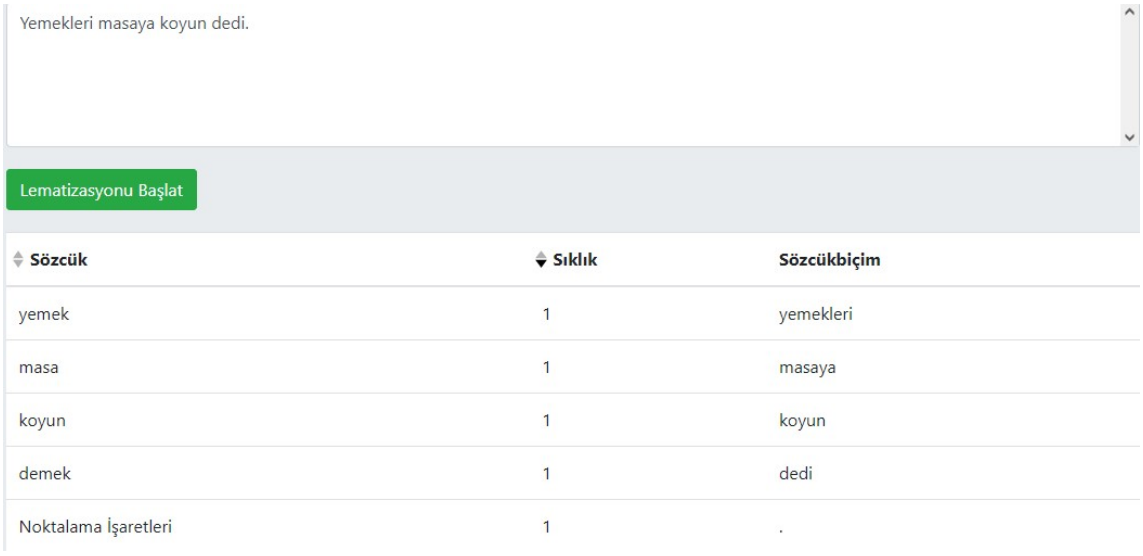



| Sözcük | Sıklık | Sözcükbiçim |
|---|---|---|
| deniz | 1 | denizlerdeki |
| koyun | 1 | koyun |
| her | 1 | her |
| iki | 1 | iki |
| yaka | 1 | yakasını |
| görmek | 1 | görmüşlerdi |
| Noktalama İşaretleri | 1 | . |

Şekil 4. "Denizlerdeki koyun her iki yakasını görmüşlerdi." cümlesinin lematizasyon sunucu.

Şekil 4'te görüldüğü gibi uygulama "koyun" sözcüğünü "koy" biçiminde lematize edememiştir. Diğer sözcükler (denizlerdeki -> deniz, yakasını -> yaka, görmüşlerdi -> görmek, her -> her, iki -> iki) doğru biçimde işlenmiştir.

Örnek cümle 2: "Yemekleri masaya **koyun** dedi." Cümlede olması gereken hedef çıktı koy-un biçiminde koymak fiilidir.




| Sözcük | Sıklık | Sözcükbiçim |
|---|---|---|
| yemek | 1 | yemekleri |
| masa | 1 | masaya |
| koyun | 1 | koyun |
| demek | 1 | dedi |
| Noktalama İşaretleri | 1 | . |

Şekil 5. "Yemekleri masaya koyun dedi." cümlesinin lematizasyon sunucu.

Şekil 5'te "koyun" sözcüğü dışındaki sözcükler doğru olarak lematize edilmiştir.

Örnek cümle 3: "İlkin **koyunun** boynuna sarılmak istedi fakat yapamadı." Hedef çıktı koyun+un biçiminde koyun ismidir.

İlkin koyunun boynuna sarılmak istedi fakat yapamadı.

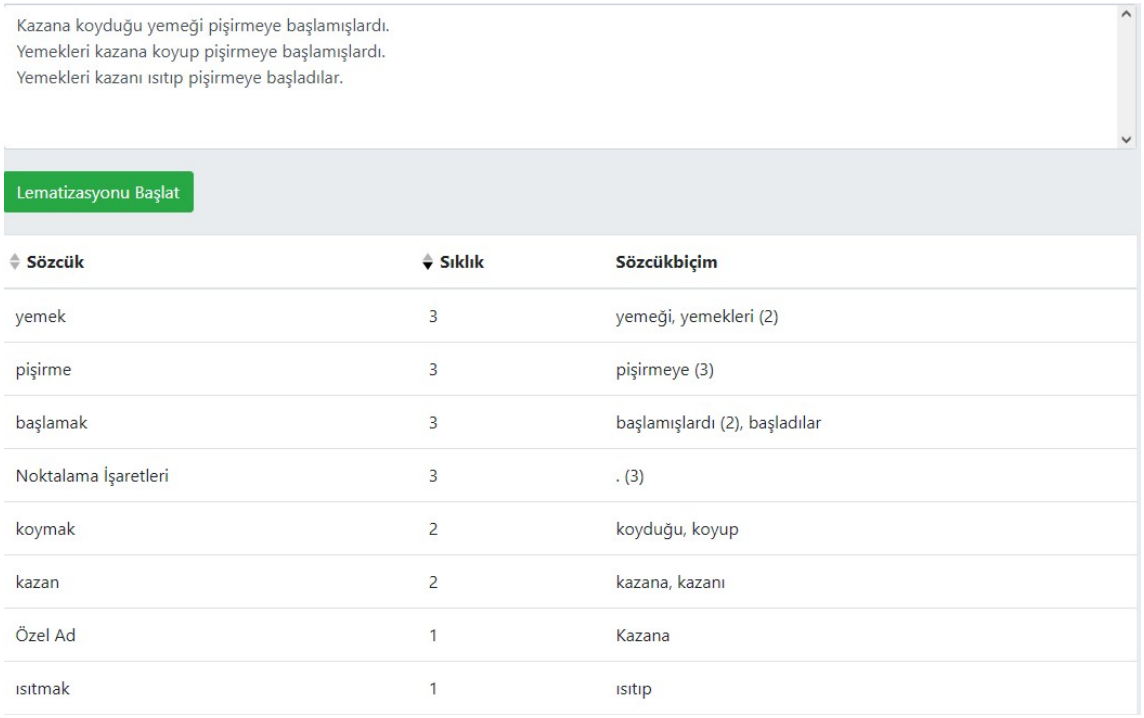
Lematizasyonu Başlat

| Sözcük | Sıklık | Sözcükbiçim |
|---|---|---|
| ilkin | 1 | ilkin |
| koyun | 1 | koyunun |
| boyun | 1 | boynuna |
| sarılmak | 1 | sarılmak |
| istemek | 1 | istedi |
| fakat | 1 | fakat |
| yapabilmek | 1 | yapamadı |
| Noktalama İşaretleri | 1 | . |

Şekil 6. İlkin koyunun boynuna sarılmak istedi fakat yapamadı cümlesinin lematizasyon sonucu.

Şekil 6'da tüm sözcükler doğru biçimde lematize edilmiştir.

Yukarıda verilen örnekler sistemi zorlayıcı örneklerdir. Sözcüklere ait bağlam sayısı arttıkça yani cümle uzunlukları ve sayıları arttıkça doğruluk oranlarının da artması beklenmektedir. Aşağıdaki örnek sonuçta "kazan" sözcüğünün içinde bulunduğu "**Kazana** koyduğu yemeği pişirmeye başlamışlardı.", Yemekleri **kazana** koyup pişirmeye başlamışlardı." ve "Yemekleri **kazanı** ısıtıp pişirmeye başladılar." cümleleri yer almaktadır.




Şekil 7. "kazan" sözcüğü lematizasyon sonuçları.

Şekil 7'de ilk cümlede yer alan "Kazana" sözcüğü uygulama tarafından hedef çıktı olan ve yemek pişirmede ya da su ısıtmada kullanılan kazan nesne ismi olarak değil özel ad olarak verilmiş, diğer "kazana" ve "kazanı" biçimleri doğru olarak sonuçlandırılmıştır.

## 4. Sonuç

DDİ'de gelinen nokta itibariyle yapay öğrenmenin dile ait birçok problemin çözümünde önemli rol oynadığı açıktır. Dünya dilleri içinde yapısı bakımından farklılık gösteren ve eklemeli dillerin tipik örneği olarak gösterilen Türkçenin biçimbilimsel açıdan zengin yapısı lematizasyon konusunda zorlu bir durum sergilemektedir. ElemanTR adıyla sunduğumuz uygulamayla %99,01'lik test başarım oranı Türkçe DDİ araştırmalarındaki bu zor problemin çözümüne katkı sağlayacak ve daha önce geliştirilen sistemlerin tasarım ve uygulamalarına yeni bir soluk getirecektir. ElemanTR adını verdiğimiz çevrim içi lematizasyon sistemi, veriye duyarlı bir sistem olduğundan eğitim verisindeki sözcüklerin sadece hikâye ve roman türlerine özgü özellikler dışında, diğer türlere dayanan daha genel özellikli derlem tabanlı çalışmalarla desteklenerek genel kapsamlı bir lematizasyon uygulamasına dönüşecektir. Sistemin veriyle desteklenmesi, oluşturulacak derlemin hem sistem başarımının artmasına hem de sonuçların günümüz Türkiye Türkçesi için farklı alanlarda madde başlarına dayanan kavram çözümleme süreçlerine fayda sağlaması olasıdır.

Sözlük hazırlayıcılar bakımından da çevrimiçi sürümü kullanıcılara sunulduğunda, derleme dayalı sözlüklerin oluşturulması sürecinde madde başlarının tespiti ve yeni sözcüklerin çıkarımında uygulamanın katkı sağlayacağı da öngörülmektedir.

**Kaynakça**

Alpaydın, E. (2011). *Yapay Öğrenme* (1. basım). Boğaziçi Üniversitesi.

Arslan, E., ve Orhan, U. (2016). Graph-based lemmatization of turkish words by using morphological similarity. In *2016 ınternational symposium on ınnovations in ıntelligent systems and applications (ınısta)*. IEEE. https://doi.org/10.1109/inista.2016.7571835

Bergmanis, T., ve Goldwater, S. Context sensitive neural lemmatization with lematus. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of* (pp. 1391–1400). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1126

Mikolov, T., Chen, K., Corrado, G., ve Dean, J. (2013, January 16). *Efficient Estimation of Word Representations in Vector Space*. https://arxiv.org/pdf/1301.3781

Müller, T., Cotterel, R., Fraser A. ve Schütze, H. (2015). *Joint Lemmatization And Morphological Tagging With Lemming*. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Öztürkmenoğlu, O., ve Alpkoçak, A. (2012). Comparison of different lemmatization approaches for information retrieval on turkish text collection. In *2012 ınternational symposium on ınnovations in ıntelligent systems and applications*. IEEE. https://doi.org/10.1109/inista.2012.6246934

Özkan, Y. (2021). *Uygulamalı Derin Öğrenme*. Papatya.