# Türkçe Wordnet'in Otomatik Oluşturulması
# Automatic Construction of Turkish Wordnet

*M.Fatih Amasyalı*

Bilgisayar Mühendisliği Bölümü, Yıldız Teknik Üniversitesi, İstanbul
mfatih@ce.yildiz.edu.tr

## Özetçe

Günümüz doğal dil işleme projelerinin özellikle kelime anlamlarını içerenlerinde Wordnet önemli bir araç olmuştur. BalkaNet projesi altında Türkçe bir Wordnet için çalışmalar başlamış fakat henüz orijinali kadar etkili kullanılabilecek bir hale gelmemiştir. Bu tür bir veri tabanının oluşturulması için işlenmesi gereken bilgi miktarı çok büyük olduğundan bu işlemin elle yapılması oldukça zahmetli ve zaman alıcı bir işlemdir. İşte bu probleme çözüm olarak bu çalışmada Türkçe Wordnet'in otomatik olarak üretilmesi için denemeler yapılmıştır. Wordnet'in ana iskeletini oluşturan "Tüm X'ler Y'dir" türündeki ilişkilerin Internet sayfalarından otomatik olarak elde edilmesinin olabilirliği araştırılmıştır. Oluşturulan sistem yaklaşık %66'lik bir doğrulukla iki kelime arasında böyle bir ilişki olup olmadığını belirleyebilmektedir. Önerilen bu metot sadece bir ilişki türüne özel olmadığından başka Wordnet ilişkileri içinde uygulanabilirdir. Sistemin daha doğru bir veritabanı üretebilmesi için uygun olarak etiketlenen kelime çiftlerinden sadece birçok kaynakta tekrar edenler veri tabanına kaydedilmelidir.

## Abstract

Wordnet is an efficient and necessary tool for recent natural language processing projects. The studies on building a Wordnet for Turkish (as a part of the Balkanet project) started. But, Turkish Wordnet is not efficient as the fifteen years old English one. Because, the manually developing process is hard and time consuming job. In this paper, 4 methods are suggested for automatic generation of some main parts of Turkish Wordnet. 2 methods are implemented and hyperonym relations (main part of Wordnet) were generated automatically with %66 success ratio. So the methods are not for only a special relation, they can be implemented for other relations. For healthier Wordnet database, only the high frequented pairs (which are repeated in different resources) must be inserted to the database.

## 1. Giriş

Kelime anlamları, doğal dil işleme projelerinde en önemli konulardan biridir. Bir kelimenin anlamı, üst sınıfı cümleleri öğelerine ayırmada, kelime anlamının açıklaştırılmasında, bilgi çıkarımında (knowledge extraction) ve benzeri birçok uygulamada kullanılan bilgilerdir.

İngilizce için Wordnet adlı etkili bir araç mevcuttur. Sistemin web ara yüzüne *http://www.cogsci.princeton.edu/cgi-bin/webwn* adresinden ulaşılabilir. Türkçe için BalkaNet projesi altında yürütülen bir çalışma vardır ancak henüz tamamlanmamış olduğundan henüz bu denli etkili bir araç değildir.

Bu çalışmada Türkçe Wordnet'in otomatik olarak oluşturulması için dört metot önerilmiştir. Raporun ikinci kısmında Türkçe Wordnet'in otomatik olarak oluşturmanın gerekliliği anlatılmış, üçüncü kısımda İngilizce için yapılan çalışmalar özetlenmiştir. Dördüncü kısımda önerilen metotlar ayrıntılı bir biçimde anlatılmış ve sonraki kısımda elde edilen sonuçlar gösterilmiştir. Sonuç bölümünde metodun kısıtları ve yorumlar yer almaktadır.

### 2. Türkçe Wordnet'i Neden Otomatik Oluşturalım?

Şu an devam eden bir çalışma varken neden Türkçe Wordnet'i otomatik oluşturalım? Şu an devam eden çalışmada, önce kısıtlı bir küme üzerinde bir çeviri yapılmıştır; ardından elle ve otomatik düzeltmeler/geliştirmeler yapılmaktadır [1]. Mart 2004 itibariyle içinde 11.628 eşküme ve 17.550 ilişki vardır. Ancak içindeki İngilizce çeviriden kaynaklanan birçok anlamı olmayan kelimenin varlığı etkili bir kullanımı zorlaştırmaktadır. Sistemin web arayüzüne *http://www.hlst.sabanciuniv.edu/TL/* adresinden ulaşılabilir.

Wordnet'in otomatik olarak olşuturulması; varolan sistemin gelişimini hızlandıracaktır ve aynı zamanda daha fazla bilgi içerebilmesine olanak sağlayacaktır. Otomatik olarak üretilmiş bilgilerle de desteklenen Türkçe Wordnet daha fazla bilgi içereceğinden daha etkili bir araca dönüşecektir.

## 3. Önceki Çalışmalar

Wordnet ilk olarak elle yapılmaya başlanmasının ardından birçok otomatikleştirme denemesi olmuştur. Bu denemelerin büyük bir kısmı belirli tür ilişkilerin otomatik olarak elde edilmesi şeklinde olmuştur. Örnek olarak Hearst,1998 (şablonlar metodunu kullanmış ve birçok ilişki türü için uygulamıştır)[2], Brin,1998 (şablonlar metodunu kullanmış ve kitap-yazar ikililerinin bulunmasında kullanmıştır)[3], Harabagiu, Miller, Moldovan,1999 (kelime anlamlarını önce lexical forma daha sonra anlamsal forma çevirmiş ve bu formlardan orijinal Wordnet'i geliştirmek için yeni bilgiler üretmiştir)[4] çalışmaları verilebilir.

## 4. Otomatik Oluşturmak için Önerilen Metotlar

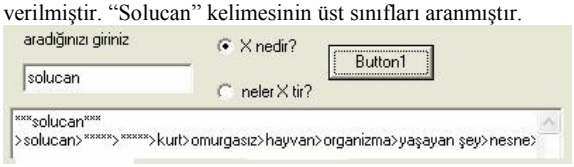
Bu bölümde önceki çalışmalardan esinlenerek önerilen 4 metot anlatılacaktır.

### 4.1. Orjinal Wordnet'ten çeviri

Sadece İngilizce Wordnet'in prolog formatındaki dosyaları ve İngilizce – Türkçe sözlük kullanarak tamamen otomatik olarak Wordnet'in Türkçeleştirilmesi için bir çalışma yapılmıştır. Çalışmanın adımları aşağıda anlatılmıştır.

0-7803-9238-8/05/$20.00 ©2005 IEEE

*1.adım:* Orijinal (İngilizce) Wordnet'in ilişkilerinin olduğu prolog dosyaları [5] Internet'ten indirildi.
*2.adım:* Her bir ilişki türü için bir veritabanı tablosu oluşturuldu ve Wordnet, Acces veritabanı formatına çevrildi.
*3.adım:* Anlamları içeren tabloda anlamlar ve bu anlamların ID'leri bulunmaktadır. Eşkümelerin her biri tablolarda bu ID'lerle tutulmaktadır. Bu özellik sayesinde sadece anlamlar tablosunun Türkçeleştirilmesi yeterli olacağı görüldü.
*4.adım:* Anlamlar tablosunun Türkçeleştirilmesi için Internet üzerinden hizmet veren bir sözlük kullanılmış ve İngilizce karşılıkları yerine sözlükteki Türkçe karşılıklarından ilki anlamlar tablosuna kaydedilmiştir. Burada ilk anlamın kullanılması Türkçe Wordnet'in doğruluğunu azaltan bir faktördür.
*5.adım:* Oluşturulan Wordnet içinde alt-üst ilişkilerini sorgulayan bir program yazıldı. Programın arayüzü aşağıda verilmiştir. "Solucan" kelimesinin üst sınıfları aranmıştır.




Şekil 1*: Otomatik çeviri ile elde edilen sistemin ara yüzü.*

Yazılan bu programla birçok denemeler yapılmış ancak İngilizce – Türkçe sözlükteki Türkçe karşılıklardan sadece ilk anlamların kullanılması ve Türkçe'de karşılığı olmayan kelimeler yüzünden programın çok iyi sonuçlar üretemediği görülmüştür. Eksiklikleri olmasına rağmen hiçbir elle müdahale gerektirmeden yapılmış olması önemli bir özelliktir.

### 4.2. Sözlük Tanımları

Aşağıda Türk Dil Kurumunun Internet üzerindeki sözlüğünde yer alan birkaç sözcüğün anlamı verilmiştir.

**kedi**: Kedigillerden, köpek dişleri iyi gelişmiş, kasları çevik ve kuvvetli evcil veya yabanî, küçük memeli *hayvan*
**timsah**: Sürüngenlerden, sıcak bölgelerin akarsularında yaşayan, kalın derili, uzun kuyruklu, iri bir *hayvan*
**kalem**: Yazmak, çizmek gibi işlerde kullanılan çeşitli biçimlerde *araç*
**cetvel**: Doğru çizgileri çizmeye yarayan, dereceli veya derecesiz, tahtadan, plastikten veya madenden yapılmış *araç*, çizgilik
**şemsiye**: Bir sapın üzerinde esnek tellere gerilmiş, açılıp kapanabilen, yağmur ve güneşten korunmak için kullanılan, su geçirmez kumaştan yapılmış taşınabilir *eşya*

Anlamlar incelendiğinde görülecektir ki anlamlarda sözcüklerin ait oldukları üst sınıflar yer almaktadır. Bu özellik sayesinde kelimelerin alt-üst sınıf ilişkileri otomatik olarak elde edilebilir. Ayrıca anlamlarda yer alan sıfatlarla da sözcüklerin karşılık geldikleri kavramların özellikleri oluşturulabilir.

Şu an mevcut Türkçe Wordnet'in oluşturulmasında[1] da sözlük tanımlarından yararlanılmıştır. Sözlük tanımlarındaki (bir tür, bir çeşit, –giller, karşıtı) gibi şablonlar kullanılarak otomatik çiftler çıkartılmıştır.

### 4.3. Şablonlar

Elde edilmek istenen ilişki içindeki örnek ikililer arasındaki sıkça rastlanan kelime ya da kelime grupları o ilişkinin şablonları olarak tanımlanır ve diğer metinlerde bu şablonlar aranır. Daha sonra bulunan şablonların sağ ve sollarındaki kelimeler arasında bu tür ilişkinin olduğu öne sürülür. Bulunan ikililerinden frekansları yüksek olanlar (çok defa farklı metinlerde geçenler) Wordnet'in veritabanına kaydedilir. Bu metodun uygulaması olan bir çalışma beşinci bölümde anlatılacaktır.

### 4.4. Öğelerine Ayrılmış Metinlerden

"ye" fiilinin nesneleri yiyecek olarak sınıflandırılabilir mantığından esinlenen bir metottur. Öncelikle metinlerin öğelerine ayrılması gerekmektedir. Daha sonra öğe ilişkileri kullanılarak kelimeler yada kelime grupları arasındaki ilişkiler otomatik olarak elde edilebilir. Şekil 2'de bu temel fikri destekleyen bir Internet sayfası yer almaktadır.


geyiq.com/forum - Taksim borsada bira ile yarım döner **yerken**
geyiq.com/forum > geyiq alanı > kıl oluyorum, dumur oldum > Taksim borsada bira ile yarım döner **yerken**. Orjinalini görmek için ...
www.geyiq.com/forum/archive/index.php/t-10219.html - 3k - Önbellek - Benzer sayfalar

**Yerken** Family Grave Search
... It's like we've always known each other!". - Pam from CA. Advertisement. Click Here. Search Page for Surname: **Yerken**. Name: First, Middle, **Yerken** Last. ...
www.findagrave.com/surnames/y/**yerken**.html - 13k - Önbellek - Benzer sayfalar

Hürriyetim
... Kelebek, 25.05.2004. Ödülü, Bush kraker **yerken** söylemeyin, ... Umarım kimse ona bu ödülü kazandığımı, o kraker **yerken** söylemez' diye yanıtladı. ...
www.hurriyetim.com.tr/ haber/0,,sid~436@nvid~416896,00.asp - 42k - Önbellek - Benzer sayfalar

MİLLİYET İNTERNET - BUSINESS
... Zeytin **yerken** alzheimer oluyoruz haberimiz yok. Zeytini, zehirli tekstil boyası ile veya demir sülfat gübresi ile karartıp satıyorlar. ...
www.milliyet.com.tr/2003/12/12/business/bus07.html - 30k - Önbellek - Benzer sayfalar

TurkiyeOnLine.com - Haber
... sağlık. Mantar **yerken** dikkat Havaların ısınmasıyla birlikte doğada ortaya çıkan mantarların bilinçsiz olarak tüketilmesinin, zehirlenmelere neden ...
www.turkiyeonline.com/haber/saglik/ haber.php?story=2004_04_02_mantar - 20k - Önbellek - Benzer sayfalar

Şekil 2*: "ye" fiilinin nesnelerinin yiyecek olarak etiketlenmesi.*

Ancak yukarıda belirtildiği gibi bu metodun kullanılabilmesi için metinlerin öğelerine ayrılması gerekmektedir. Bu işlem ise çok kelime içeren cümleler için çok basit bir işlem olmadığından bu metotla belki az kelimeli cümleleri kullanarak otomatik ilişki çıkarımı gerçekleştirilebilir. Bu metodun uygulanması konusunda çalışmalar devam etmektedir.

## 5. Alt-Üst sınıf ilişkileri için Deneysel Bir Çalışma

Bu bölümde şablonlar metodunun bir uygulaması anlatılacaktır. Bu uygulamada Wordnet'in omurgasını oluşturan alt-üst sınıf ilişkilerine odaklanılmıştır ancak başka ilişki türleri içinde uygulanabilir bir metot geliştirilmiştir.

Uygulamada önce alt-üst sınıf ilişkisini içeren örnek ikililer bulunmuş(kedi-hayvan, tornavida-araç, Türkiye-ülke gibi) ve bu ikililer Google'da aratılıp aralarında kalan kelime yada kelime grupları(şablonlar) bulunmuştur. Aşağıda bulunan şablonların sıkça rastlananlarından örnekler verilmiştir.

| | | |
|---|---|---|
| ve diğer | ve benzeri | türü olan |
| ler ve diğer | veya diğer | lar ve diğer |

Bir sonraki adımda Google'ın sonuç sayfalarında ve Türkçe 400.000 web sitesinin kaynak kodunu içeren büyük bir metin arşivinde bu şablonlar aranmıştır. Bulunduğu yerin sağ ve solundaki kelimeler kelime sonuna kadar alınıp

kaydedilmiştir. Buradaki ara amaç kelimeleri değil bu tür ilişkiyi içeren "kelime+ek"leri belirlemektir.

Bulunan ikililer geçtikleri cümledeki halleriyle [6] adresindeki araç kullanılarak eklerine ayrılmıştır. Aracın birden fazla sonuç ürettiği durumlarda ilk çözüm kabul edilmiştir. Bu işlemlerin ardından elde edilen ikililere örnek aşağıda verilmiştir.

- **dua**lar ve her türlü **ibadet**
  - Noun+ A3pl+ Pnon+ Nom (dualar)
  - Noun+ A3sg+ Pnon+ Nom (ibadet)

Internet'teki araç kullanılarak eklerine ve köklerine ayrılan verilerde 53 farklı kelime türü ve ek olduğu görülmüştür. Bu durumda bir örnek için (53*2)+şablon(1) olmak üzere 107 adet özellik elde edilmiştir. Buradaki "kelime türü ve ek" özellikleri o ek veya kök türü varsa 1 yoksa 0'dır. Diğer bir ifadeyle her bir "kelime1+ek+şablon+kelime2+ek" şeklindeki örnek 1'ler ve 0'lardan oluşan 107 uzunluğunda bir dizi olarak ifade edilmektedir.

Yukarıda anlatılan şekilde 730 adet örnek oluşturulmuştur. Asıl amacımız bu ilişki türüne sahip kelime1+ek+şablon+kelime2+ek ifadelerini öğrenmektir. Bu işlemi gerçekleştirmenin 2 yolu vardır. İlki bu işlemi elle yapmak ikincisi ise bir sınıflandırma metodu kullanmak. Bu çalışmada ikincisi tercih edilmiştir.

Bir sınıflandırma metodu kullanmak için öncelikle eğitim ve test verileri gerekmektedir. Burada elde edilen 730 örneğin yarısı eğitim yarısı test için ayrılmıştır ve eğitim ve test verilerinin tamamı elle uygun olup olmadıklarına göre etiketlenmiştir. Bu sayede 2 sınıf elde edilmiş ve herbir örneğin hangi sınıfa ait olduğu işaretlenmiştir.

### 5.1. Özellik Seçimi

Sınıflandırma yaparken bir örnekten alınan tüm özelliklerin kullanmak yerine bazılarını seçip kullanmak sınıflandırma başarısını nasıl etkiler? Bu sorunun cevabını bulmak için Tablo 1'deki örneği inceleyelim.

*Tablo 1*: Örnek Eğitim Kümesi

| 1. Özellik | 2. Özellik | Sınıf |
|---|---|---|
| 1 | 3 | A |
| 2 | 3 | B |
| 1 | 4 | A |
| 2 | 3 | B |

Tablo 1'deki 1.özellik sınıfları belirlemede 2.özelliğe göre daha uygundur. Böylesi küçük bir tabloda bunu gözle görebilmek mümkünken daha fazla özellik ve örnek içeren tablolarda(ki bizim çalışmamızda bu tablonun 107 özellik+ 1 sınıf etiketi=108 adet sütunu, 730/2=365 adet-eğitim seti için-satırı vardır) bu işlemi gerçekleştirmek için çeşitli metotlar geliştirilmiştir. Bu çalışmada bu metotlardan "Sinyalin Gürültüye Oranı(Signal to Noise Ratio) ve Temel Bileşenlar Analizi(Principal Component Analysis) kullanılmıştır.

Sinyalin Gürültüye Oranı metodunda herbir özellik için Eşitlik 1'de verilmiş olan s değeri hesaplanır.

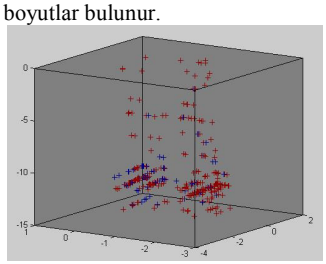
$$s_i = \frac{\mu_{c1} - \mu_{c2}}{\sigma_{c1} - \sigma_{c2}} \qquad (1)$$

$\mu_{c1}$ → c1 sınıfındaki örneklerin i. özelliklerinin ortalaması

$\sigma_{c1}$ → c1 sınıfındaki örneklerin i. özelliklerinin standart sapması

Metodun dayandığı temel özellik, sınıflar arası ayrılıkların fazla; sınıf içi ayrılıkların az olmasının sınıflandırma başarısını arttırmasıdır. S değeri en yüksek olan özellikler sınıflandırmada kullanılırlar.

Özellik seçimi metotlarından bir diğeri de Temel Bileşen Analizidir. Bu metotta örneklerin en fazla değişim gösterdikleri boyutlar bulunur.

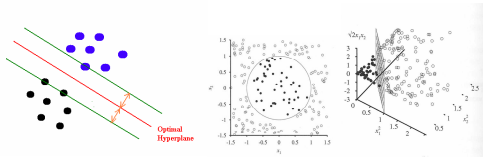


Şekil 4*: Eğitim verilerinin 3 boyuttaki gösterimi*

Şekil 4'te 107 boyutlu eğitim verilerimizin 3 boyutlu uzaya iz düşürülmüş halleri bulunmaktadır. Görüldüğü gibi örneklerin sınıfları birbiri içine geçmiş durumdadır ve bu sınıflandırmanın çok kolay yapılamayacağının sinyalidir.

### 5.2. Sınıflandırma Metodları ve Performans Ölçümü

Bu çalışmada Çok Katmanlı Algılayıcı (Multi Layer Perceptron-MLP), Öğrenmeli Vektör Kuantalama (Learning Vector Quantization-LVQ) , Destek Vektör Makineleri (Support Vector Machines-SVM) olmak üzere 3 adet sınıflandırıcı kullanılmıştır. Çalışmanın bu bölümünde kullanılan sınıflandırıcılardan Support Vector Mechine (SVM) anlatılacaktır. Vapnik[8] tarafından geliştirilen bu metot günümüzde performansı sayesinde oldukça popüler olmuş bir metottur. Sınıfları birbirinden ayıran özel düz bir çizginin bulunmasını amaçlar. Şekil 5.a'daki durumda bu çizginin bir çok çizilebilme ihtimali vardır. SVM bu çizgilerden her iki sınıfa en uzak olanını bulur. Bu sayede hataya tolerans en fazla hale gelecektir.




Şekil 5(a,b,c)*: Destek vektör makimeleri ile sınıflandırma.*

Lineer olarak ayrılamayan örnekler için örnekler daha yüksek boyutlu başka bir uzaya taşınır ve sınıflandırma o uzayda yapılır. Örneğin Şekil 5.b'de örnekler iki boyutlu uzayda $[x_1,x_2]$lineer(doğrusal) olarak ayrılamazken üç boyutlu uzayda $[x_1^2,x_2^2, \sqrt{(2x_1x_2)}]$ (Şekil 5.c) ayrılabilmektedirler.

Sınıflardaki örnek sayılarının birbirinden farklı olduğu durumlarda sınıflandırıcı performansları genelde Eşitlik 2'deki gibi değil Eşitlik 3'teki şekilde ölçülür. Örneğin bir sınıfta 90 diğer sınıfta 10 örnek bulunan bir veri setimiz olsun. Sınıflandırıcımız 100 örneğin hepsine birinci sınıf derse Eşitlik 1'e göre %90'lık bir başarı elde etmiş olur. Ancak bu yanıltıcı bir sonuçtur. Performans Eşitlik 3'teki gibi ölçülürse %45 olarak bulunur.

$$Performans(\%) = \frac{dss}{\ddot{o}s} \qquad (2)$$

$$Performans(\%) = \frac{1}{k}\sum_{i}^{k}\frac{dss(i)}{\ddot{o}s(i)} \qquad (3)$$

Eşitlik 2 ve Eşitlik 3'te;

dss→ tüm örneklerin kaçının doğru olarak sınıflandırıldığı
ös→ toplam örnek sayısı
k→ sınıf sayısı
dss(i)→i.sınıftaki örneklerin kaçının doğru olarak sınıflandırıldığı
ös(i) → i. sınıftaki örnek sayısını ifade etmektedir.

Tablo 2'de üç farklı sınıflandırma metodunun, 2 farklı boyut azaltma tekniğiyle birlikte kullanıldığında elde edilen başarı oranları verilmiştir. Başarı oranları hesaplanırken Eşitlik 3 kullanılmıştır. Test kümemizde 1. sınıf için 262, 2. sınıf için 103 olmak üzere 365 adet örnek vardır. Yüzdelik başarı oranları bu rakamlar üzerinden hesaplanmıştır.

*Tablo 2*: Farklı sınıflandırıcılar ve farklı boyut azaltma teknikleri için yüzdelik başarı oranları.

| Boyut azaltma metotları | Sınıflandırma Metotları | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | MLP | | | LVQ | | | SVM | | |
| | 1. | 2. | T | 1. | 2. | T | 1. | 2. | T |
| Orijinal boyut | 100 | 0 | 50 | 100 | 0 | 50 | 94 | 18 | 56 |
| PCA 2 | 73 | 55 | 64 | 69 | 54 | 62 | 100 | 0 | 50 |
| PCA 3 | 84 | 42 | 63 | 68 | 54 | 62 | 100 | 0 | 50 |
| PCA 5 | 77 | 37 | 57 | 75 | 51 | 63 | 100 | 0 | 50 |
| **PCA 7** | **88** | **45** | **66** | 75 | 50 | 63 | 100 | 0 | 50 |
| PCA 10 | 100 | 0 | 50 | 74 | 51 | 63 | 100 | 0 | 50 |
| S2N 2 | 100 | 0 | 50 | 100 | 0 | 50 | 100 | 0 | 50 |
| S2N 3 | 79 | 38 | 58 | 80 | 38 | 59 | 100 | 0 | 50 |
| S2N 5 | 84 | 44 | 64 | 84 | 44 | 64 | 84 | 44 | 64 |
| **S2N 7** | **88** | **45** | **66** | 83 | 44 | 64 | **88** | **44** | **66** |
| S2N 10 | 74 | 47 | 61 | 77 | 44 | 61 | 80 | 42 | 61 |

Tablo 2'de her sınıflandırıcının iki sınıftaki yüzdelik başarı oranları ayrı ayrı verilmiştir. Ayrıca 'T' ile gösterilen sütunda Eşitlik 3 ile hesaplanan başarı oranı verilmiştir. Orijinal boyutta 107 boyutlu verilerle sınıflandırma yapılmıştır. PCA ve S2N satırlarının yanlarındaki rakamlar, verinin kaç boyuta indirgendiğini göstermektedir.

Tablo 2'de görüleceği gibi en yüksek sınıflandırma başarısına(%66) 7 boyuta indirgenmiş verilerle MLP ve SVM erişmiştir. Boyut azaltmanın performansa olumlu bir yönde katkısı olmuştur.

Geliştirilen sistem sayesinde aralarında belirlediğimiz bir şablon olan iki kelimenin istediğimiz ilişki içinde olup olmadığının tahmini %66 doğrulukla otomatik olarak yapılabilmektedir. Tablo 3'te otomatik olarak bulunan ikililerden örnekler verilmiştir.

*Tablo 3*: Otomatik olarak bulunan kelime çiftlerinden örnekler.

| **Alt Sınıf** | **Üst Sınıf** | **Alt Sınıf** | **Üst Sınıf** | **Alt Sınıf** | **Üst Sınıf** |
|---|---|---|---|---|---|
| çimento | madde | cüzzam | hastalık | konut | yapı |
| karaciğer | sakatat | insan | canlı | Kars | il |
| akciğer | organ | kitap | kaynak | otobüs | taşıt |
| kalay | madde | traktör | araç | Çin | ülke |
| kafa | uzuv | kene | parazit | sayman | yetkili |

### 5.3. Frekansların Kullanımı

Sınıflandırma başarısı %100 olmadığından eğer sistem bu haliyle kullanılırsa sistemdeki verilerin %34'ü yanlış olacaktır. Bu nedenle daha doğru bir veritabanı için bulunan ikililerin frekanslarının kullanımı önerilmektedir. Bulunan ikililerden sınıflandırıcının doğru olarak etiketlediği örnekler alınır. Her bir örneğin farklı metinlerde kaçar kez geçtiği bulunur. Türkçe Wordnet veritabanına sadece belirli bir eşik değerinden daha fazla geçenler kaydedilir. Ancak bunun yapılabilmesi için çok büyük metin arşivleri işlenmek zorundadır. Bu sayede daha sağlıklı bir veritabanı elde edilebilir.

## 6. Sonuç

Bu çalışmada Türkçe Wordnet'in otomatik olarak oluşturulması için önerilen 4 metottan (otomatik çeviri, sözlük tanımlarının kullanılması, şablonlar, öğelerin kullanılması) 2'si (otomatik çeviri, şablon) uygulanmış ve sonuçlar alınmıştır. Çeviri metodu için Orijinal Wordnet ilişkileri kaydedilmiş ve Türkçeleştirilmiştir. Daha sonra bu veritabanını alt-üst ilişkileri için sorgulayan bir program yazılmıştır. Ancak Türkçe'de karşılığı olmayan kelimeler yüzünden ve çok anlamlı kelimeler yüzünden programın çok iyi sonuçlar üretemediği görülmüştür. Bu kısıtlarına rağmen büyüklüğü ve tamamen otomatik olarak üretilmiş olması avantajları olarak görülebilir.

Uygulanan ikinci metot ise şablonlar metodudur. Bu metotta her bir ilişki için özel şablonlar bulunmuş ve bu şablonlar etrafındaki kelimelerin o şablonun gösterdiği ilişki içinde oldukları öne sürülmüştür. Elde edilen örnekler eklerine ayrılarak ekler ve şablonlar uzayında ifade edilmiştir. Daha sonra örnekler istenilen ilişkiye sahip olup olmadıklarına göre elle 2 sınıfa ayrılmışlardır. Bu sayede problem iki sınıflı bir sınıflandırma işlemi haline getirilmiştir. Sistemde birçok sınıflandırma metodu ve boyut azaltma tekniği kullanılmış ve en yüksek sonuç MLP ve SVM metodlarıyla %66 olarak elde edilmiştir. Ancak bu başarı oranının yeterli olmadığı düşülmüş ve işleme bulunan ikililerin frekanslarının da katılması önerilmiştir.

Sonuç olarak kısa zamanda gerçekleştirilen yazılımlarla Türkçe Wordnet'in otomatik olarak oluşturulmasının belirli ilişki türleri için mümkün olduğu gösterilmiştir. Sistemin eksiklikleri olarak birden fazla kelimeden oluşan kelime grupları arasındaki ilişkilerin bulunamaması ve çok büyük metin dosyalarının işlenmesine gerek duyması söylenebilir.

Gelecekteki çalışmalar içim kelime gruplarının arasındaki ilişkilerin bulunması ve öğelerine ayrılmış metinlerin kullanılması düşünülmektedir. Ayrıca sistemin başarı oranının arttırılması için kelimelerin ve eklerin uzayda farklı şekillerde ifade edilmesi de araştırılabilir.

## 7. Kaynakça


[1] Building a Wordnet for Turkish, Orhan Bilgin, Özlem Çetinoğlu, Kemal Oflazer, Romanian Journal of Information Science and Technology, volume 7, 2004

[2] M. A. Hearst. Automated discovery of WordNet relations. In C. Fellbaum, editor, WordNet: an Electronic Lexical Database. MIT Press, 1998

[3] http://maya.cs.depaul.edu/~classes/ect584/papers/brin.pdf

[4] http://acl.ldc.upenn.edu/W/W99/W99-0501.pdf

[5] http://www.cogsci.princeton.edu/2.0/

[6] http://fens.sabanciuniv.edu/TL/cgi-bin/mymorp_keyb.cgi

[7] http://www.cs.sfu.ca/cs/people/GradStudents/fchena/personal/Content%20Analysis%20of%20Video%20Using%20Principal%20Components.ppt

[8] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1995